UNIVERSIDADE DA CORUÑA

FACULTADE DE INFORMÁTICA

Departamento de Computación

PhD Thesis

Automatic grading of ocular hyperaemia using image processing techniques

María Luisa Sánchez Brea

2017

PhD advisors:

Noelia Barreira Rodríguez

Antonio Mosquera González

To my family and friends

Acknowledgements

It has been a while since I started my PhD, almost four whole years. Looking back, it has been stressful, time-consuming, maddening... and also a wonderful experience and a decision that I will never regret taking. Discovering research has been discovering a fulfilling world, full of challenging and interesting projects. I know that I am in debt with a lot of people who helped me reach this point with their support and their advice, so I will try my best to thank everyone here.

I want to start by sending a huge thank you to my family and friends. They may not be there helping with my research, but I could not have made it this far without their support, both in the good and, specially, the bad times. So, I want to start by thanking my parents, my uncle and aunt (and my cousins, of course!), and my grandparents. Thank you for being there for me, I am so lucky to have you! Next, of course, I want to thank all the friends that have been by my side, enduring my madness. Thank you so much for that, for all the good times and laughs. Don't ever change!

Now, on a more serious note, I would like to express my gratitude to all the people that helped me through this work. First, I would like to thank my PhD advisors, Noelia and Antonio, for all their advice and their patience with my mistakes. I also want to extend these thanks to all the colleagues from VARPA group, that also helped me with their knowledge or support (and the occasional rant about all that is going wrong is a really valuable support!). I would also like to thank Noelia Sánchez Maroño (University of A Coruña, LIDIA group) for her help with the feature selection experiments and her valuable advice on the matter. As this is an interdisciplinary work, it could not have been possible without the collaboration of the Optometry Service (University of Santiago de Compostela). Specially, many thanks to Carlos and Hugo for labelling the

images that I needed and providing clinical insight.

Finally, I would like to thank Katharine Evans from the School of Optometry and Vision Sciences (Cardiff University) for a wonderful internship, and for her additional help with the clinical part of the work. And, of course, all the PhD students from the School. You made me feel at home and among good friends since the first day. Thank you all!

"It is important that we know where we come from, because if you do not know where you come from, then you don't know where you are, and if you don't know where you are, you don't know where you're going. And if you don't know where you're going, you're probably going wrong."

Terry Pratchett

Abstract

The human eye is affected by a number of high-prevalence pathologies, such as Dry Eye Syndrome or allergic conjunctivitis. One of the symptoms that these health problems have in common is the occurrence of hyperaemia in the bulbar conjunctiva, as a consequence of blood vessels getting clogged. The blood is trapped in the affected area and some visible signs, such an increase in the redness of the area, appear.

This work proposes an automatic methodology for bulbar hyperaemia grading based on image processing and machine learning techniques. The methodology receives a video as input, chooses the best frame of the sequence, isolates the conjunctiva, computes several image features and, finally, transforms these features to the ranges that optometrists use to evaluate the parameter. Moreover, several tests have been conducted in order to analyse how the methodology reacts to unfavourable situations. The goal was to cover some common issues that assisted diagnosis methodologies have to face in real-world environments.

The proposed methodology achieves a significant reduction of the time that the specialists have to invest in the evaluation. Thus, it has a direct repercussion on reaching a fast diagnosis. Moreover, it removes the inherent subjectivity of the manual process and ensures its repeatability. As a consequence, the experts can gain insight in the parameters that influence hyperaemia evaluation.

Keywords

Computer aided diagnosis, Optometry, hyperaemia, Image segmentation, Feature selection, Regression

Resumen

El ojo humano se ve afectado por un gran número de patologías de alta prevalencia, tales como el Síndrome del Ojo Seco o la conjuntivitis alérgica. Uno de los síntomas que estos problemas de salud comparten es la aparición de hiperemia en la conjuntiva bulbar, consecuencia del taponamiento de vasos sanguíneos. La sangre queda atrapada en el área afectada y aparecen signos visibles, como el aumento de rojez en la zona.

Este trabajo propone una metodología automática para la evaluación de hiperemia bulbar basada en técnicas de procesado de imagen y aprendizaje máquina. La metodología recibe un vídeo, escoge la mejor imagen de la secuencia, aísla la conjuntiva, calcula varias características en la imagen y, por último, transforma estas características al rango de valores que los optometristas usan para evaluar la hiperemia. Además, se han realizado varias pruebas para analizar como reacciona la metodología a situaciones desfavorables. El objetivo era incluir problemas comunes que aparecen a la hora de aplicar una metodología de asistencia al diagnóstico en un entorno real.

La metodología propuesta consigue una reducción significativa del tiempo que los especialistas invierten en la evaluación. Por lo tanto, tiene repercusiones directas en alcanzar un diagnóstico rápido. Además, elimina la subjetividad inherente al proceso manual y garantiza su repetitibilidad. Como consecuencia, los expertos pueden obtener información acerca de los parámetros involucrados en la evaluación de la hiperemia.

Palabras clave

Diagnóstico asistido por ordenador, Optometría, Hiperemia, Segmentación de imágenes, Selección de características, Regresión

Resumo

O ollo humano vese afectado por un elevado número de patoloxías de alta prevalencia, tales como o Síndrome do Ollo Seco ou a conxuntivite alérxica. Un dos síntomas que ditos problemas de saúde teñen en común é a aparición de hiperemia na conxuntiva bulbar, consecuencia da conxestión dos vasos sanguíneos. O sangue queda atrapado na área afectada, e aparecen signos visibles, como o incremento do arrubiamento na zona.

Este traballo propón unha metodoloxia automática para a avaliación da hiperemia bulbar baseada en técnicas de procesado de imaxe e aprendizaxe máquina. A metodoloxía recibe un vídeo como entrada, escolle a mellor imaxe da secuencia, illa a conxuntiva, calcula varias características da imaxe e, por último, transforma estas características ós rangos que os optometristas usan para avaliar o parámetro. Ademáis, leváronse a cabo varias probas para analizar como reacciona a metodoloxía ante situacións pouco favorables. O obxectivo era abarcar algúns dos problemas máis comúns que atopan as metodoloxías de asistencia á diagnose en entornos reais.

A metodoloxía proposta consegue unha redución significativa do tempo que os especialistas invirten na avaliación. Polo tanto, ten unha repercusión directa na obtención dunha diagnose rápida. Ademáis, elimina a subxectividade inerente ó proceso manual, e asegura a súa repetitibilidade. Como consecuencia, os expertos poden entender mellor os parámetros que influencian a avaliación da hiperemia.

Palabras clave

Diagnóstico asistido por ordenador, Optometría, Hiperemia, Segmentación de imaxes, Selección de características, Regresión

Contents

List of Figures

List of Tables

Part I

Hyperaemia

Chapter 1

The eye and its pathologies

The eye is one of the most complex organs in the human body, and not without a reason. In a small area, several structures are interconnected and work in perfect harmony in order to obtain a clear representation of our environment. However, they are delicate structures, and can be affected by a large number of pathologies, some of them directly, and some of them as a previous symptom of more serious issues. Because of this, the eyes pose a high relevance in medical diagnosis.

The anatomy of the human eye, seen from above, is depicted in Fig. 1.1. The cornea, the clear front surface of the eye, focuses the light into the eye. The iris regulates the amount of light that reaches the back of the eye by adjusting the size of the pupil. The crystalline lens, located behind the pupil, helps to further focus the light. The retina is a light-sensitive layer located in the back of the eye that receives light and creates electrical impulses in response. These impulses travel through the optic nerve until they reach the visual cortex in the brain. The retina is covered by a vascular layer, the choroid, and the centre of the eye is filled with a jelly-like substance, the vitreous. The outer part of the eye is the sclera, an opaque protective layer that is totally white. Finally, there is an additional layer that covers the sclera: the conjunctiva. This work is centred in that external layer, whose main function is to protect the sclera. Specifically, the primary focus of the work is the occurrence of hyperaemia in this part of the conjunctiva, a parameter that serves as a starting point for the diagnosis of several pathologies.

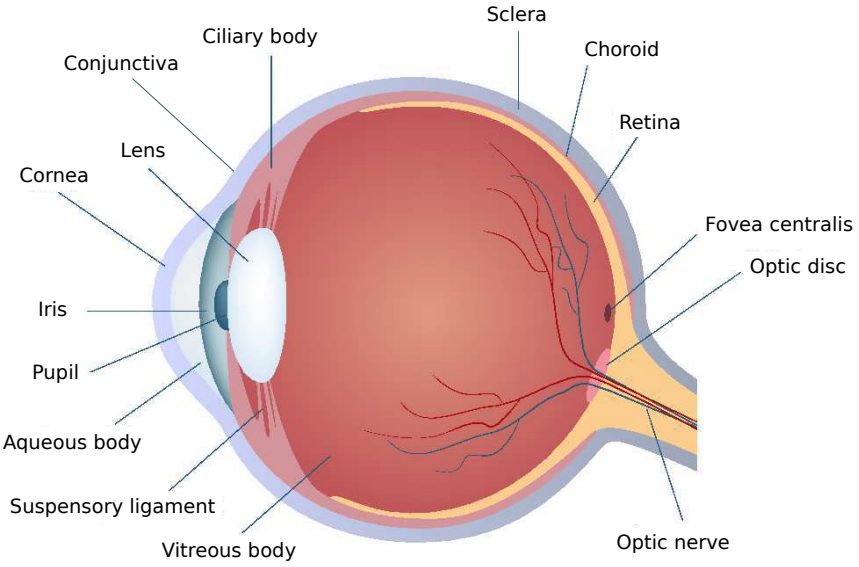


Figure 1.1: Anatomy of a human eye.

1.1 The conjunctiva

The conjunctiva is a formation of layers of flattened epithelial cells arranged upon a cutaneous membrane. It is highly vascularised and covers from the sclera to the inside of the eyelids, as depicted in Fig. 1.2. To main parts can be observed:

Tarsal conjunctiva also known as palpebral conjunctiva, is the part that covers the inside of the eyelids (Fig. 1.2, right).

Bulbar conjunctiva also known as ocular conjunctiva, is the part that protects the sclera. The epithelium is loosely attached to the sclera and moves with it (Fig. 1.2, left).

The area where both the bulbar and tarsal conjunctivas join is called fornix conjunctiva, and covers the junction between eyeball and eyelids.

The main purpose of the conjunctiva is to help with the lubrication of the eye, as it has its own means of secreting tears and mucus, although in a smaller quantity than the lacrimal gland. It also serves as a protective barrier to the eye, limiting the entrance of microbes.

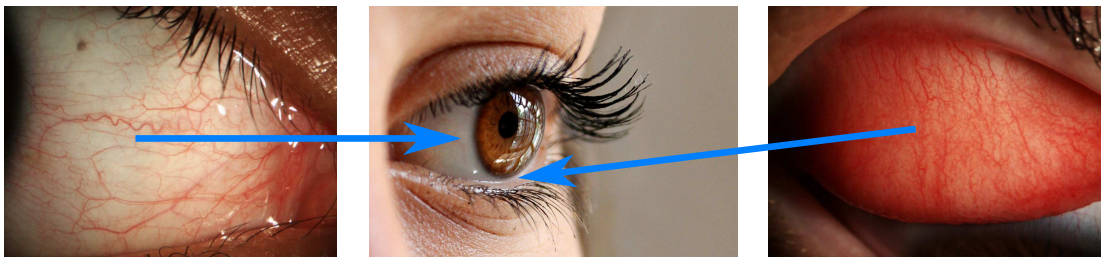


Figure 1.2: Left and right: close-up view of the bulbar and tarsal conjunctivas. Centre: the situation of each conjunctiva within the eye.

1.2 Common pathologies

The conjunctiva is a common cause of medical consultation, as it is exposed to external agents and highly susceptible of irritation, allergies, dryness or infections among other issues. Some of its most common problems are the following:

Allergic conjunctivitis [1] is a reaction that occurs in the conjunctiva when exposed to an allergen, usually pollen or spores. As the conjunctiva becomes irritated, the red colouration appears in the area. Allergies are one of the most common pathologies, affecting from 10 to 20% of the world's population. Even though not every patient that has an allergy will develop conjunctivitis, they have a higher risk. Conjunctivitis is associated with several uncomfortable symptoms, such as itchiness, pain or burning sensation. In the most severe cases, these symptoms can grow in intensity, so the patient may present sensibility to light or even vision loss.

Subconjunctival haemorrhages [1] appear when a vessel breaks, creating a red colouration due to the lost blood that has yet to be absorbed by the conjunctiva. They may have several causes, from the presence of foreign elements touching the surface of the eye to a severe bout of coughing. Usually, a single subconjunctival haemorrhage is not an indicator of an underlying problem. However, if they appear frequently they must be monitored, as they are a common symptom of diabetes, hypertension or blood disorders.

Dry eye syndrome [2, 3] appears when the eye is not correctly lubricated. This can happen due to the low quality or low quantity of the tears. Its causes are multiple, from environmental conditions, such as pollution, to low blink rate, such as long periods working in front of a computer. It displays a high incidence, affecting between 5 and 14% of the world's population [4], and it has a growing tendency. Hence, it is considered a public health problem. It causes discomfort and pain, and can even lead to corneal damage if not treated properly.

Use of contact lenses [2, 5], specially a prolonged exposure, can cause irritation on the eye. Moreover, it can be the cause of associated issues, such as the appearance of dry eye syndrome. The symptoms can vary from slight discomfort to acute pain, and so can the associated problems. Contact lenses can cause corneal abrasion, eye infections or even corneal ulcers.

Certain medications have been confirmed to alter the appearance of the conjunctiva. One of the most common examples are the topical medications used to treat elevated intra-ocular pressure[6]. However, medications that are not in direct contact with the conjunctiva can also show side-effects, such as anticoagulants.

Glaucoma is a group of diseases. There are three main types: primary open-angle, angle-closure and normal-tension, and the most common one, primary open-angle glaucoma, is hereditary. Glaucoma is caused by an abnormal rise of the intra-ocular pressure, that damages the optic nerve and can lead to permanent vision loss [3, 7]. The risk of developing a glaucoma increases with age and, due to the absence of early warning signs and the potentially severe consequences, regular checkups are advised.

Besides the aforementioned, there are other less common pathologies that present some of their early symptoms in the conjunctiva, such as diabetes, blepharitis, corneal abrasion, keratitis, iritis or scleritis.

1.3 Clinical tests

In order to diagnose the aforementioned pathologies, several clinical tests can be tackled. These tests analyse the surface of the eye by looking at a certain symptom, with or without the help of chemical agents or specific instruments. The following procedures [8] are some of the most common, and are performed by optometrists in primary care:

Tear meniscus height. The tear meniscus is an accumulation of tears in the lower lid margin, and it is useful to estimate the tear volume. This parameter can be measured from several media sources, such as optical coherence tomography in cross-section or slit-lamp microscope [9].

Break-up time (BUT). This test measures the interval between the last blink of the patient and the moment when his/her tear film starts to disappear (break). To that end, fluorescein is instilled (Fig. 1.3, left).

Non-invasive break-up time (NIBUT). As fluorescein can influence the tear film, a non-invasive version of the test was developed. While BUT can be assessed with a video camera, NIBUT needs to be measured with a special device, such as a Keratometer, a hand-held Keratoscope or a Tearscope.

Ocular hyperaemia. This parameter can be measured in both the tarsal or bulbar conjunctivas. The concept of hyperaemia includes different changes that the conjunctiva overcomes as a consequence of vessel engorgement in the tissue. It is a non-invasive test, based on the observation of the conjunctiva.

Phenol red thread test (PRTT). A cotton thread is placed under the patient's conjunctival fornix. The thread is dyed in phenol red dye, a product that is pH sensitive and, hence, it changes colours when in presence of tears.

Corneal and conjunctival staining. The staining indicates a disruption of the epithelium, caused by damaged cells. The most common approaches to measure the staining on the surface of the eye are the application of a lissamine green or fluorescein dye. These elements allow to see the points where the epithelium is disrupted (Fig. 1.3, middle).

Lid-wiper epitheliopathy (LWE). The staining of the upper and lower lid margin regions can happen as a consequence of the contact with the ocular surface or with contact lenses wearing, caused by an increase in friction [10]. It is commonly assessed by applying lissamine green and fluorescein (Fig. 1.3, right).

Ocular Surface Disease Index. The formulation of a series of questions, where each of the possible answers has an associate score, is a helpful test to assess the symptoms of a patient. It is used when there are hints pointing at the existence of Dry Eye Syndrome [11].

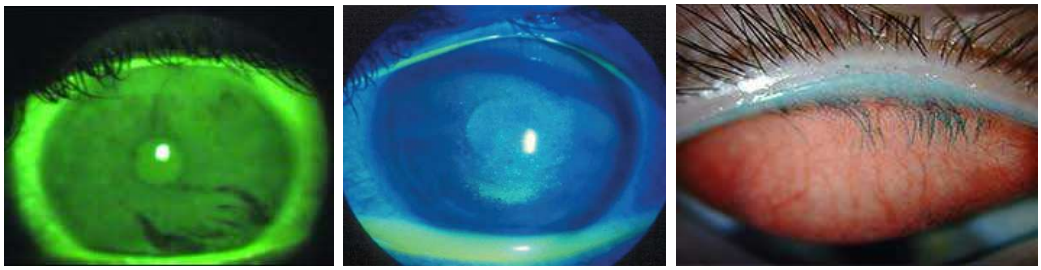


Figure 1.3: Some of the most common medical trials performed in the conjunctiva. From left to right: BUT test (the black areas show the rupture of the tear film), conjunctival staining test (the whiter points in the middle represent the stains) and lid-wiper epitheliopathy test.

As each of the aforementioned tests is focused in one parameter at a time, the specialists usually run more than one of them in order to diagnose a pathology.

1.4 Bulbar hyperaemia

The term hyperaemia represents the engorgement of the blood vessels in a tissue, which causes an increase of blood in the area. It can appear due to normal bodily processes, or as a sign of medical problems. One of the commonly affected tissues is the ocular conjunctiva. Both bulbar and tarsal conjunctivas can be affected, and to analyse each area can help to diagnose different diseases. To assess the tarsal conjunctiva is more uncomfortable for the patient, as the eyelid has to be turned inside out, while the bulbar hyperaemia can be evaluated with direct observation and without touching the patients' eye. Moreover, larger image databases exist in bulbar conjunctiva.

Bulbar hyperaemia is a medical condition related to several of the most common diseases that affect to the conjunctiva, such as allergic conjunctivitis [12] or dry eye syndrome [8]. The most characteristic visual sign is the occurrence of a red colouration in the sclera, caused by the blood vessel engorgement. This condition, known as erythema, is one of the first symptoms that appear in an unhealthy conjunctiva. Figure 1.4 depicts eyes with different grades of hyperaemia.

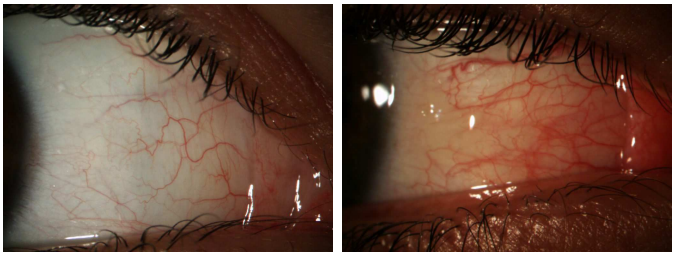


Figure 1.4: Example of eyes that present different levels of bulbar hyperaemia.

Chapter 2

Evaluation of hyperaemia

Bulbar hyperaemia is an early symptom of common pathologies. However, the current evaluation that clinicians have to perform has several drawbacks as it is subjective, non repeatable and tedious for the optotmetrists. In this chapter, this manual evaluation of hyperaemia is described. Then, the datasets that are used to develop the automatic methodology are presented. Moreover, an additional effort is made in analysing the available data, both the videos or images that optometrists use to grade and the assigned evaluations themselves.

2.1 Hyperaemia assessment in clinical practice

Conjunctival hyperaemia is diagnosed by direct observation of the patient's eye with a slit lamp. However, it is a common practice to capture the patient's eye in some kind of digital media, in order to allow the further analysis of the case or the exchange of information among several clinicians (Fig. 2.1). Therefore, the manual process starts by recording a video or taking photographs of the patient's eye. In the case of filming a video, more information is stored, and the specialist has the opportunity to decide the best depiction of the conjunctiva from a wide spectrum of frames. In the case of taking several pictures, the specialist can capture or record different orientations of the eye and focus on the most relevant areas.

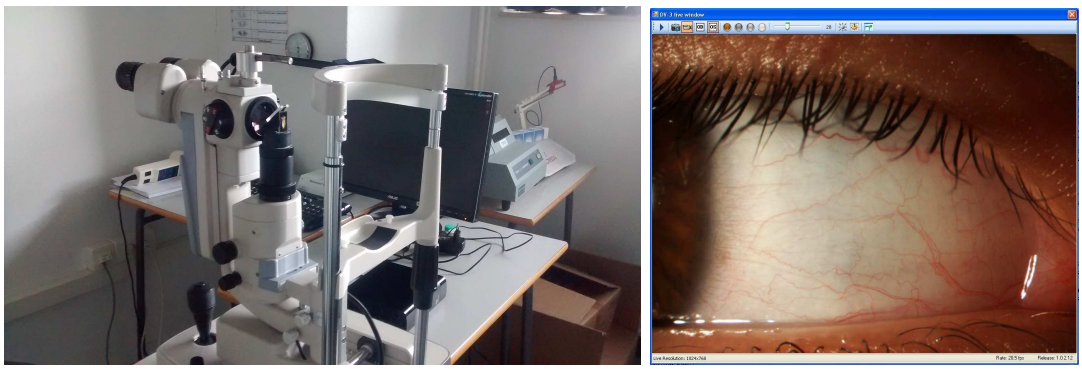


Figure 2.1: Manual capture procedure. Left: Topcon DV-3 digital camera attached to the slit-lamp. Right: example screenshot of the Topcon IMAGEnet i-base software during a normal hyperaemia recording.

When the starting point is a video, the optometrist must analyse the sequence in order to find the frame that offers the best depiction of the conjunctiva. Next, the clinician analyses this image, looking at indicators such as vessel quantity or hue of the background. The occurrence or not of these features and their interactions will determine the severity of the symptom by comparing them with a given grading scale. A grading scale is a collection of sorted images, drawings or photographs, that establish levels of severity against with each patient's eye can be compared and assigned a grade [5].

There are several scales that have been developed to assess bulbar hyperaemia. The first known scale for evaluating bulbar hyperaemia was proposed in 1987 by McMonnies and Chapman-Davies (MC-D), and is depicted in Fig. 2.2. Since then, a large number of scales have been proposed. Two of the most widely-used are the Efron and the BHVI (Brien Holden Vision Institute, formerly known as CCLRU, Cornea and Contact Lens Research Unit) grading scales. The former consists of a set of five drawings that depict the conjunctival redness, ranging from 0 to 4 (Fig. 2.3, top). The latter consists of a set of four photographs, and ranges from 1 to 4 (Fig. 2.3, bottom). In these scales, a lower value indicates a more normal clinical symptom level. We can observe how different scales may depict slightly different areas of the conjunctiva, which could influence the evaluation [13].

Figure 2.2: MC-D scale for bulbar conjunctival redness grading.

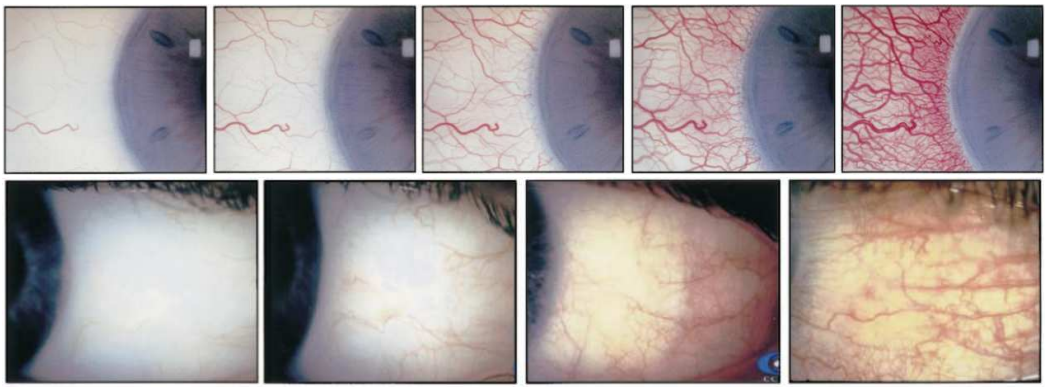


Figure 2.3: Efron and BHVI scales for bulbar conjunctival redness grading.

Although the scales consist in a small, finite set of prototypes, the evaluation of hyperaemia in practice is performed with a higher precision. Since there is a wide spectrum of cases between each two levels of the scale, using decimal points helps to improve the representation of the state of the patient. Thus, the optometrists usually use a real number expressed with one decimal value in order to represent the proximity from the patient to the closest prototype. Therefore, the values can be seen as a continuous range rather than individual classes. However, there are exceptions to this rule, as some scales have a wider range of represented values, such as Validated Bulbar Redness (VBR), with range 0-100 (Fig. 2.4).

Figure 2.4: VBR scale for bulbar conjunctival redness grading. From left to right: values 10, 30, 50, 70 and 90.

Nevertheless, the manual process has several drawbacks. First, it presents a high subjectivity, both intra- and inter-expert, and is non-repeatable. This subjectivity ap-

pears among the different steps of the process. As there are no clear indicators or objective goodness metrics, several sources of bias are stacked together. Besides, the procedures to obtain the videos or images are not standard, so there is noticeable variance in illumination, focus and distance to the camera. Additionally, the environmental conditions and the equipment may also vary. Also, the specialists' behaviour regarding which image features are taken into account is highly subjective and heavily influenced by their past experiences. As a consequence, the image characteristics that are involved in the process are difficult to define. Blinks or movements of the patients can also hinder the process. Finally, the manual procedure is tedious and time-consuming, specially if the initial media is a video, as the optometrist has to review the whole sequence. The relevance of this issue increases if the specialist has to review several patients in a row, which is a common situation.

Therefore, there is a clear need for the automatisation of the process. This can be achieved by means of computer vision and machine learning techniques through several steps. Thus, the first step is to gain insight on the problem that is being tackled. In this particular scenario, that implies to know the necessary optometry details as well as the technical details that may be useful for the implementation. However, an important step that is sometimes overlooked is to study the inputs themselves, that is, to analyse the distribution of the available data and the existing underlying relationships and, ideally, to discover what causes both of them.

2.2 Analysis of hyperaemia assessments

The objective of this study is to depict the special characteristics of the data, in order to gain a better understanding of the decisions taken during the development of the methodology.

First, the scale categories are discrete values, but the specialists rate the images using decimal values. However, they do not use all the values within the scale range, as differences in the images below a certain threshold are not appreciable for the human eye. One decimal place is generally employed, but some experts prefer to use only a half integer step between evaluations, or a quarter of integer. Some studies [14, 15]

conclude that it comes down to personal preference, experience and how confident the expert is in the grading. Therefore, even though the evaluations are continuous values in practice, the experts' evaluations can be represented as discrete classes. For example, considering decimal precision, there would be a total of 41 classes for the Efron scale (41 intervals of 0.1 from 0.0 to 4.0) and 31 for the BHVI scale (31 intervals of 0.1 from 1.0 to 4.0).

Second, the features that specialists look at in the conjunctiva are not clearly defined. An example of qualitative description of hyperaemia in each level of the Efron scale (Fig. 2.3, top) is the following [16]:

Grade 0: both the conjunctiva and the limbus are white, with one major vessel at most. The cornea is clear or there is a small white corneal reflex.

Grade 1: small increase in conjunctival and limbal redness, with the major vessel more engorged and a slight increase in number (one or two). The cornea is clear or there is a small white corneal reflex.

Grade 2: there is a further increase in conjunctival and limbal redness. The major vessels are more engorged and they slightly increase in number with slight ciliary flush. A small white reflex can be seen on the cornea.

Grade 3: conjunctiva and limbus are very red. The major vessels are engorged and marked, with a ciliary flush along the whole area. A speckled corneal reflex can be observed.

Grade 4: conjunctiva and limbus are extremely red. The major vessels are engorged, marked and some of them also show a superficial reflex. There is an intense ciliary flush all over the conjunctiva and limbus. A hazy corneal reflex can be observed.

Besides, it is common that different specialists assign different importance to different parameters. Moreover, optometrists are usually unable to explain clearly which are the most important features, as they subconsciously consider others. Sometimes a parameter is only relevant when it appears simultaneously with others, but the information regarding these interactions is difficult to model. This creates a problem, as

it is common that, even if a specialist manages to perform a successful assessment of the hyperaemia level, the knowledge applied is nearly impossible to communicate and model. Therefore, the automation has an additional advantage in the research of the underlying causes of the symptom, which can potentially lead to the discovering of new information.

Finally, it must be noted that extreme values of bulbar hyperaemia are very uncommon. Most healthy individuals present at least traces of the parameter, so a completely white eye is a rare sight. A severe case is also unlikely to be captured, as the symptom is usually tackled before it reaches the highest level. Therefore, the data sets are imbalanced, as most of the patients present intermediate levels of hyperaemia, while there are few samples of extreme cases. Moreover, the scales themselves favour a certain degree of imbalance, as the degree of change between two contiguous prototypes varies from the lowest to the highest grades. This influences also the perception of the specialists that evaluate the data sets, as the inherent subjectivity of the process causes them to adapt their evaluations to more closely resemble their previous assessments.

2.3 Description of the data sets

Two data sets were used in this thesis. The first one, VID, consists of videos, while the second one, IMG, consists of images. Most of the patients depicted were caucasian. The datasets present some visual differences, and hence some variation is expected. Their particularities are detailed in the following sections.

2.3.1 Video data set (VID)

This dataset consists of 163 videos of the bulbar conjunctiva recorded at the Optometry Service of the Faculty of Optometry (University of Santiago de Compostela, Spain). The procedure was reviewed by the University of Santiago de Compostela Ethics Committee, and followed the tenets of the Declaration of Helsinki. The patients were informed of the protocol and signed a consent form.

The videos were captured with a slit lamp camera (Topcon DV-3, Oakland, NJ). They are about 20 seconds long, with a framerate of 7fps. The resolution of each frame

is 1024×768 pixels. The structure of the videos is similar: they start with black frames and receive progressive illumination up to a peak. Then, it decreases again. In some videos, the last frames of the sequence are also black. These changes in illumination are controlled by the specialist, who modifies the intensity of the light source manually. Bulbar hyperemia evaluation requires a bright illumination, but to maintain that level of brightness for a long time is uncomfortable for the patient. Thus, the low illumination at the beginning allows the optometrist to calibrate the camera and give indications to the patient. Fig. 2.5 depicts an example of video illumination through time.

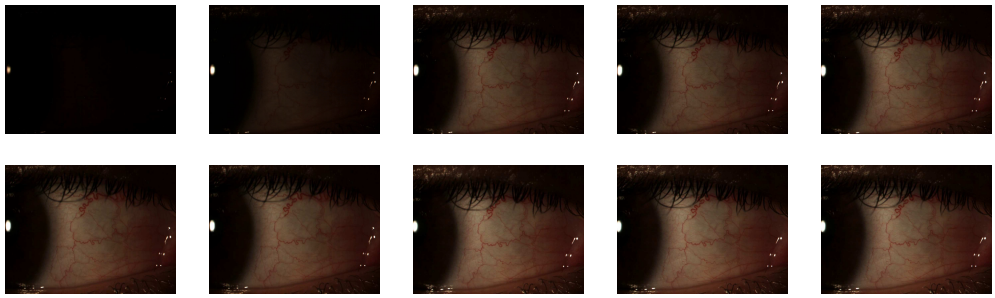


Figure 2.5: Ten frames from a hyperaemia video at different points of a video sequence from VID dataset. The top left frame was taken at the second 1.4 of the video (tenth frame of the video), and the subsequent frames are also separated 1.4 seconds (10 frames).

The videos show a side view of the patient's eye. There are four possible videos that can be taken of each patient, two of each eye. One is taken with eye looking temporally (towards right for the right eye and left for the left eye), which allows image capture of the nasal bulbar conjunctiva. The other is taken with the eyes looking nasally, so that the temporal conjunctiva can be captured. There is not information regarding the patient associated to each video, nor all the four videos were necessarily captured for all the patients.

All videos have been graded manually by two experts. They assigned a value using both the Efron and the BHVI scales, to each previously selected frame. The specialists did not exchange information during the evaluation, nor they were aware of each patient's identity. A second evaluation was performed on the same video and the same frame months apart by the same two specialists. The precision of these four evaluations was one decimal position.

Additionally, a third expert performed another grading using only the Efron scale. Again, this specialist did not have any knowledge on the patients' identities nor any additional information on the images. The precision of this evaluation was 0.2.

Figure 2.6 depicts the distribution of the evaluations of each specialist, with the values divided with a step of half point in the scale. For the specialists that performed two evaluations, the mean value is represented.

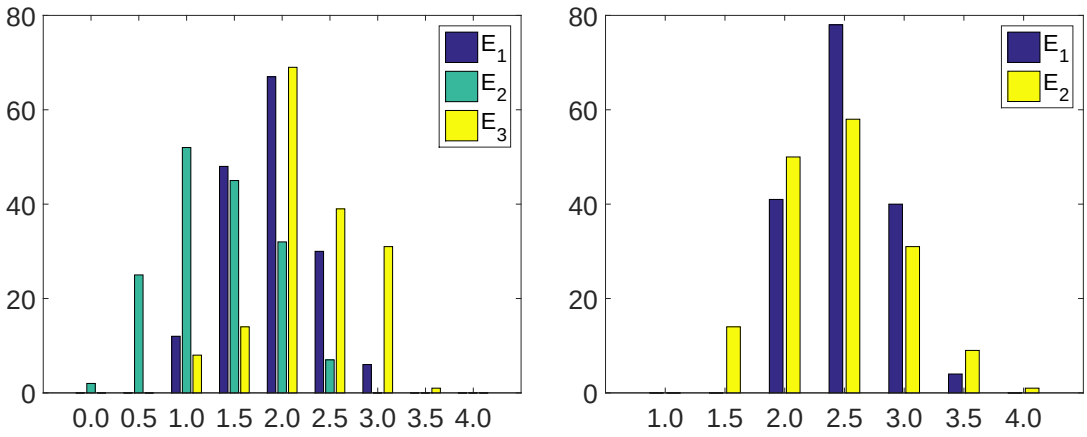


Figure 2.6: Distribution of the VID data set evaluations. Left: values for the Efron scale. Right: values for the BHVI scale.

2.3.2 Image data set (IMG)

This dataset consists of 915 images of the bulbar conjunctiva obtained at the School of Optometry and Vision Sciences (Cardiff University, Wales) as part of an study regarding contact lenses comfort[1]. The procedure was reviewed by the School of Optometry and Vision Sciences Research Audit Ethics Committee, and followed the tenets of the Declaration of Helsinki. The patients were informed of the protocol and signed a consent form.

The images were captured with a slit lamp camera (Bon 75-SL DigiPro3 HD, Bonn, Germany). The image resolution is 1600×1200 pixels. Fig. 2.7 depicts an example of images from the data set.

[1] http://research.cardiff.ac.uk/converis/portal/Project/2525952

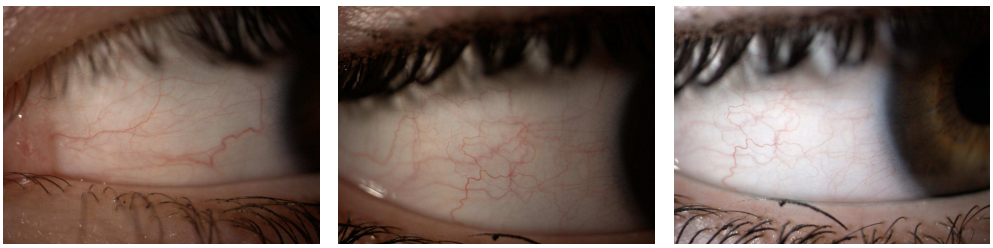


Figure 2.7: Images from the IMG data set.

The images show a side view of the patient's eye, and belong to both eyes and both sides of the eye, from the iris side to the caruncle (the small, reddish nodule at the inner corner of the eye) or the corner of the eye side. The images belong to 35 patients. Each patient went through four checkups. In the first one, used as baseline, the eye of the patient is depicted without external agents. Then, the patients were asked to wear their contact lenses during two weeks in a row, and then received the second checkup, that consists of images of the eyes while wearing contact lenses. The third checkup took place after a non-wearing contact lenses (washout) period of 7 days, and depicts the patients' eyes without contact lenses. Finally, the fourth checkup took place after another two week period of wearing contact lenses, and depicts once more the patients' eyes while wearing contact lenses. In all the cases, the contact lenses were worn on a daily basis for an average of 10 hours. Moreover, some of the images, independently of the checkup, depict the eye with remains of a blue dye used to detect conjunctival staining. At each checkup, four image types were taken, these were left eye, nasal side (LEN); left eye, temporal side (LET); right eye, nasal side (REN); and right eye, temporal side (RET) (Fig. 2.8). All the images have a similar disposition, and one or more images were obtained for each type in a given checkup.

The whole dataset was graded manually by an optometrist, who divided each image in areas (upper and bottom halves, and left and right halves) and evaluated each area separately using the Efron scale with steps of 0.25. The specialist had access to the knowledge of the patients' identities and previous checkups. Figure 2.9 shows the distribution of the evaluations of the specialist when the values were divided with a step of half point in the scale.

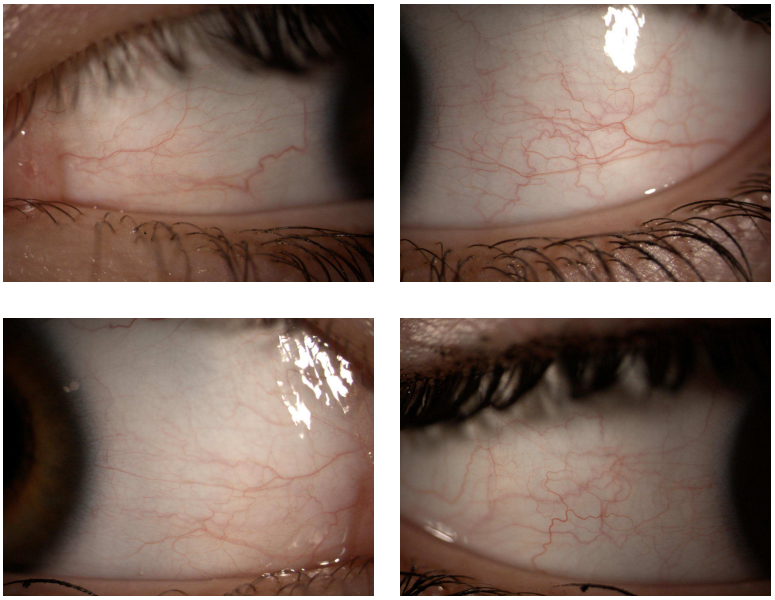


Figure 2.8: Different eyes and sides for a certain patient and checkup. From left to right and top to bottom: LEN, LET, REN and RET.

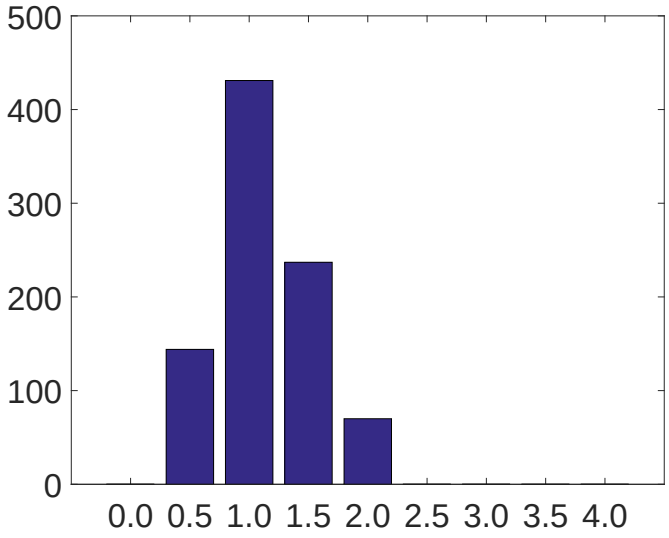


Figure 2.9: Distribution of the IMG data set evaluations.

2.4 Analysis of the experts' evaluations

In order to gain a better understanding on the problem, the first step is to analyse and compare the experts' evaluations in each dataset. To that end, the correlation and kappa index among the evaluations of the same image have been computed.

2.4.1 Correlation and kappa index in the VID dataset

As it was mentioned in the introduction, the evaluation of hyperaemia in the bulbar conjunctiva is a highly subjective task. The expert's evaluations vary widely depending on variables such as the moment when the evaluation takes place or the previous experiences. Figure 2.10 illustrates this intra-expert variability, with each axis of the plot representing one of the two evaluations of an expert (performed months apart) in the VID set. The x-axis depicts the first evaluation, and the y-axis, the second.

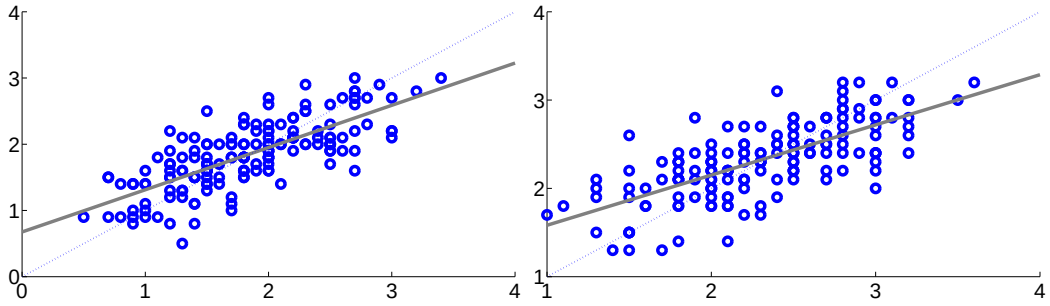


Figure 2.10: Intra-expert variability for the VID set. Each axis represents one of the two evaluations of the same expert. Left: values for the Efron scale. Right: values for the BHVI scale.

Moreover, Figure 2.11 depicts the differences among several experts grading the same patient. The inter-expert differences are also large, and, therefore, the mean between the two evaluations that each specialist performed was computed, and these values, compared. This helps to establish a more accurate comparison, removing some of the intra-expert variability. Even then, the variability continues to be noticeable.

As the variability in the data can be misleading to the machine learning algorithms, those images where the experts differ more than a given threshold were removed. The distribution achieved after a threshold of 0.5 is depicted in Fig. 2.12. The inter-expert

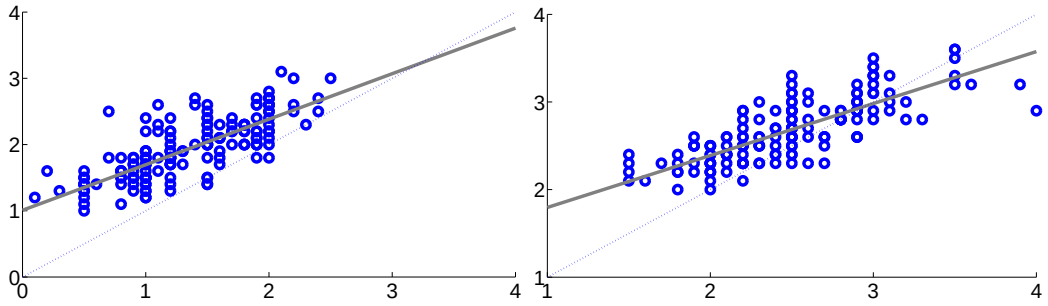


Figure 2.11: Inter-expert variability for the VID set. Each axis represents the evaluations of one of the experts. Left: values for the Efron scale. Right: values for the BHVI scale.

correlation rises from 0.69 to 0.83 and from 0.59 to 0.65 in the Efron and the BHVI scales, respectively. 114 images fulfil the 0.5 restriction in both scales.

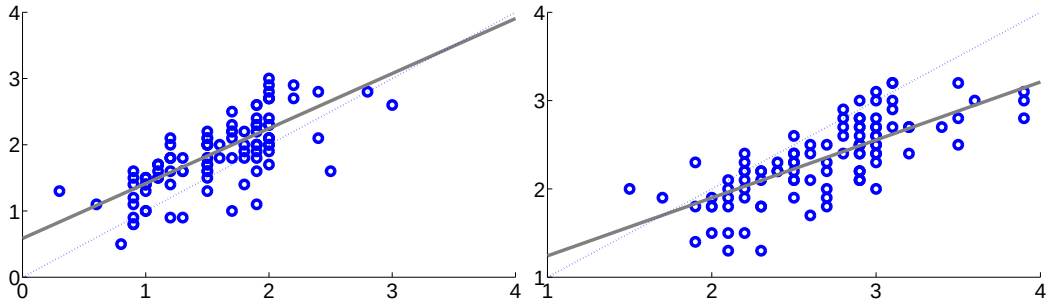


Figure 2.12: Inter-expert variability for the VID_1 image set. Each axis represents the evaluations of one of the experts. Left: values for the Efron scale. Right: values for the BHVI scale.

Additionally, the Cohen's kappa coefficient was calculated in each of the three aforementioned situations. As this is a regression problem, the data has to be transformed into discrete classes in order to compute this value. Therefore, the experts' evaluations were divided into discrete partitions with steps 0.1, 0.5 and 1.0 by assigning each value to the closest prototype. Thinner partitions increase the number of classes, lowering the agreement. The values for each experiment are depicted in Tables 2.1, 2.2 and 2.3. In these tables, *po* and *pe* represent the observed and random agreement, respectively. The null hypothesis H_0 is that the observed agreement is accidental. The significance level α is 0.05. The agreement is displayed in a scale from 0 to 5, from lower to higher (poor, slight, fair, moderate, substantial, high). Further information on the kappa in-

dex can be found in Appendix D. It can be observed how the agreement is higher in the BHVI scale, but the best results only achieve moderate agreement, and only with the wider divisions.

Table 2.1: Cohen's kappa coefficient for the evaluations of two experts in the VID dataset.

Efron scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0245	0.0415	-0.0177	0	0.0003	0.3331	Accept
0.5	0.1840	0.1878	-0.0046	0	0.0028	0.9307	Accept
1.0	0.4540	0.3956	0.0965	1	0.0082	0.2870	Accept
BHVI scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.1350	0.0651	0.0748	1	0.0005	0.0005	Reject
0.5	0.5092	0.2866	0.3120	2	0.0031	0.0000	Reject
1.0	0.7362	0.4736	0.4989	3	0.0070	0.0000	Reject

Table 2.2: Cohen's kappa coefficient for the evaluations of two experts in the VID_1 dataset.

Efron scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0857	0.0512	0.0364	1	0.0006	0.1399	Accept
0.5	0.3238	0.2536	0.0941	1	0.0035	0.1135	Accept
1.0	0.6381	0.4782	0.3064	2	0.0072	0.0003	Reject
BHVI scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0476	0.0532	-0.0058	0	0.0007	0.8198	Accept
0.5	0.3429	0.2583	0.1140	1	0.0050	0.1069	Accept
1.0	0.6190	0.4272	0.3349	2	0.0121	0.0023	Reject

Taking into account the results, two subsets of the VID dataset were chosen in order to perform certain experiments. First, VID_1 dataset, that considers only the evaluations of the two specialists that graded the set twice. This set includes the 114 images where one expert (considering the average of his/her evaluations) differs from the other in less than 0.5. The average of the four evaluations is taken as ground truth. The objective of this subset is to raise the agreement between the experts, so that the machine learning techniques have a better performance. Next, a second subset of 50 videos was randomly selected. The videos were tagged by two specialists, who selected the best frame. Two manual segmentations of the conjunctiva were performed for each

Table 2.3: Cohen's kappa coefficient for two evaluations of the same expert in the VID dataset.

Efron scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.1104	0.0524	0.0612	1	0.0003	0.0010	Reject
0.5	0.4356	0.2488	0.2487	2	0.0019	0.0000	Reject
1.0	0.6687	0.4365	0.4121	3	0.0042	0.0000	Reject
BHVI scale							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.1411	0.0606	0.0857	1	0.0004	0.0000	Reject
0.5	0.5583	0.2881	0.3795	2	0.0024	0.0000	Reject
1.0	0.7055	0.4758	0.4383	3	0.0052	0.0000	Reject

of these 50 optimal frames. This reduced subset was selected because all the videos from VID dataset follow a similar structure and, thus, a smaller subset is representative enough. Table 2.4 summarises the features of each subset of VID dataset.

Table 2.4: Summary of refined datasets with VID_n origin.

Name	# elements	# evaluations	Selection rules
VID_1	114	4	$E_1(x) - E_2(x) \leq 0.5, x \in VID$
VID_2	50	2	Random selection

2.4.2 Correlation and kappa index in the IMG dataset

The IMG dataset has only one associated evaluation by a single optometrist. Therefore, the intra- or inter-expert subjectivity cannot be evaluated. However, 141 of the images of the dataset were manually chosen, so they did not present blue dye or contact lenses, nor did they present image issues such as blurriness. Another two specialists graded these images, also using the Efron scale. They did not exchange information during the evaluation nor were they aware of each patient's identity. The precision of these two evaluations was one decimal position.

This dataset, labelled as IMG_1, can be analysed regarding inter-expert subjectivity. Figure 2.13 depicts the differences among several experts grading the same image. Each plot depicts the comparison of a pair of experts. The correlation values are 0.338, 0.333 and 0.661, respectively. It can be observed how E_2 and E_3 gradings are similar, while

E_1 differs strongly. The distribution of values of the two experts that agree the most is depicted in Fig. 2.14.

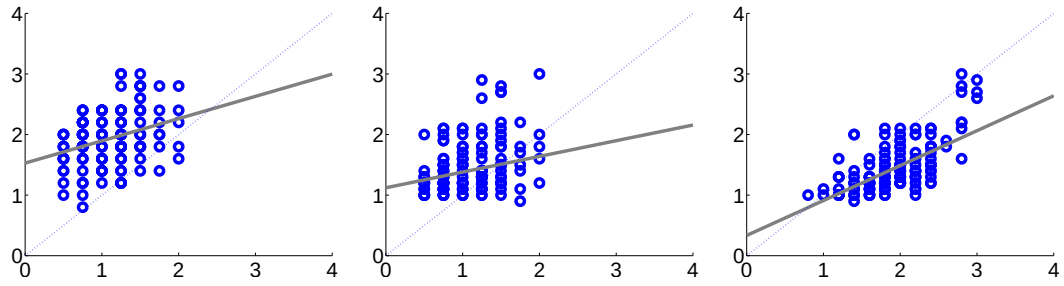


Figure 2.13: Inter-expert variability for the IMG_1 set. Each axis represents the evaluations of one of the experts. From left to right: E_1 vs E_2, E_1 vs E_3 and E_2 vs E_3.

By following the same idea as with the VID dataset, the images where the experts differ more than a given threshold were removed. The distribution achieved after the threshold of 0.5 is depicted in Fig. 2.15. The inter-expert correlation for each pair rises to 0.812, 0.688 and 0.898, respectively. The image set is reduced to 39, 86 and 76 images, respectively.

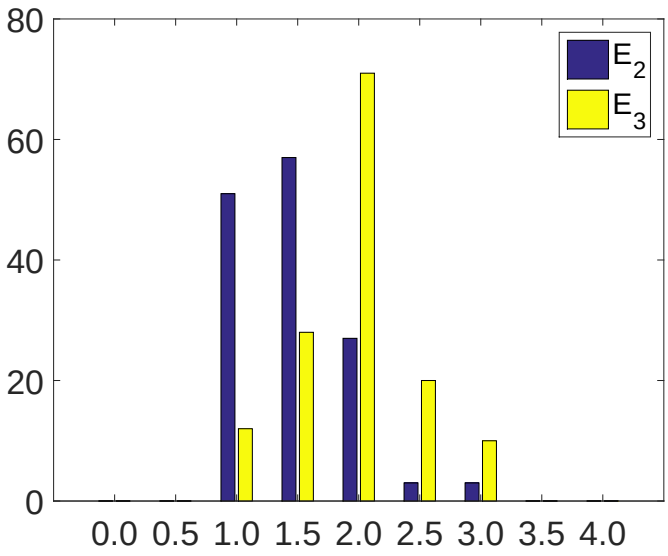


Figure 2.14: Distribution of the IMG_1 data set evaluations for experts E_2 and E_3.

The 76 images where the experts' evaluations from E_2 and E_3 differ less than 0.5 points was labelled as IMG'_1. This reduced image set has a correlation of almost 0.9.

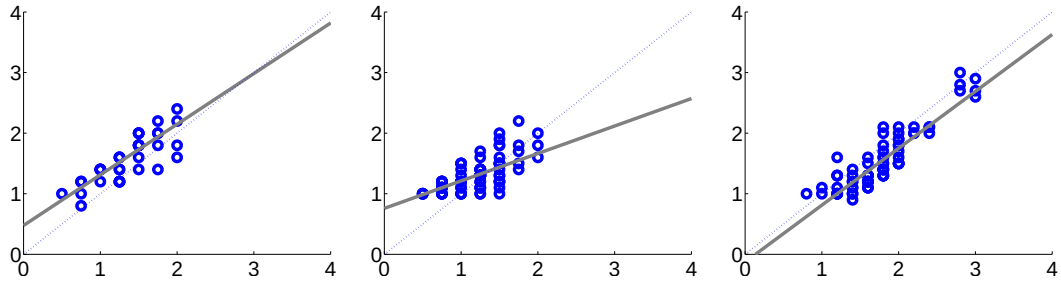


Figure 2.15: Inter-expert variability for the IMG'_1 set. Each axis represents the evaluations of one of the experts. From left to right: E_1 vs E_2, E_1 vs E_3 and E_2 vs E_3.

The ground truth for the machine learning algorithms is the average value of the two evaluations.

The kappa index for both IMG_1 and IMG'_1 is depicted in Tables 2.5 and 2.6. The main difference with VID evaluations is the substantial agreement shown in Table 2.6 for the comparison of E_1 and E_2, consequent with the results observed in Fig. 2.15.

Table 2.5: Cohen's kappa coefficient for the evaluations of two experts in IMG_1.

E_1 vs E_2							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0142	0.0212	-0.0071	0	0.0009	0.8103	Accept
0.5	0.1348	0.1565	-0.0258	0	0.0046	0.7026	Accept
1.0	0.3121	0.3051	0.0101	1	0.0186	0.9411	Accept
E_1 vs E_3							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.1418	0.0697	0.0775	1	0.0009	0.0083	Reject
0.5	0.4043	0.3279	0.1136	1	0.0033	0.0476	Reject
1.0	0.6383	0.5393	0.2149	2	0.0082	0.0179	Reject
E_2 vs E_3							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0426	0.0500	-0.0078	0	0.0007	0.7590	Accept
0.5	0.2199	0.2120	0.0100	1	0.0044	0.8804	Accept
1.0	0.4823	0.3727	0.1746	1	0.0136	0.1337	Accept

In order to summarise, the subsets that were considered from IMG dataset are depicted in Table 2.7. First, for the sake of a more precise analysis, the IMG_1 set was selected, as the manually selected images present the optimal conditions to hyperaemia grading. Moreover, three evaluations are available for this subset. Then, the subset of

Table 2.6: Cohen's kappa coefficient for the evaluations of two experts in IMG'_1.

E_1 vs E_2							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0513	0.0460	0.0055	1	0.0040	0.9301	Accept
0.5	0.4872	0.3176	0.2486	2	0.0140	0.0355	Reject
1.0	0.8718	0.4997	0.7438	4	0.0258	0.0000	Reject
E_1 vs E_3							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.2326	0.0995	0.1477	1	0.0019	0.0008	Reject
0.5	0.6628	0.3940	0.4436	3	0.0078	0.0000	Reject
1.0	0.8372	0.6071	0.5857	3	0.0121	0.0000	Reject
E_2 vs E_3							
step	po	pe	kappa	agreement	var	p	H_0
0.1	0.0789	0.0538	0.0265	1	0.0013	0.4557	Accept
0.5	0.4079	0.2630	0.1966	1	0.0064	0.0143	Reject
1.0	0.7237	0.4062	0.5347	3	0.0154	0.0000	Reject

IMG'_1 was also selected, as experts E_2 and E_3 have a high correlation in these images, hence being more adequate to use as ground truth for machine learning algorithms. Finally, in order to study the repeatability of each stage of the methodology, 20 pairs of images were selected. Each pair of images belongs to the same eye and the same side. One of the images of the pair show the eye in optimal conditions to grade hyperaemia, while the other shows the eye with some type of *alteration*. Thus, ten of the pairs show the same image with and without remains of a blue dye (S_{blue}), while the other ten pairs show the same image with and without contact lenses (S_{cont}). The remaining 875 images of the dataset were grouped in IMG_2.

Table 2.7: Summary of refined datasets with IMG_n origin.

Name	# elements	# evaluations	Selection rules
IMG_1	141	3	Manual selection
IMG'_1	76	2	$E_2(x) - E_3(x) \leq 0.5, x \in IMG_1$
IMG_2	875	1	$IMG - S_{blue} - S_{cont}$
S_{blue}	20	1	2 images of 10 eyes, with and without blue dye
S_{cont}	20	1	2 images of 10 eyes, with and without contact lenses

2.5 Discussion

In IMG data set there are images of the same eye of the same patient at different checkups. Moreover, there are several images showing different regions of the same eyes. One can obtain several frames from each video of VID data set, but the variability is much lower than in the different areas of IMG dataset, as the eye of the patient is usually static.

VID data set only shows the naked eye, without any alterations, while some of the images in IMG were captured while the patient was wearing contact lenses or with remains of dye in the surface of the eye.

Most of the images from IMG dataset barely show any eyelids, while most of VID data set shows a large part of them. This is because the camera is closer to the patient. However, in a few images the situation is the opposite, with the camera positioned further from the patient's eye, as depicted in Fig. 2.16. This is specially relevant for the segmentation of the region of interest, as the shape and size of the conjunctiva differs from one dataset to the other. The distance to the camera also influences the skin colour, since in the images where the eye is closer to the camera, the conjunctiva hue is similar to the skin tone. As a consequence, the skin colour is even closer to the conjunctiva hue in IMG data set (Fig. 2.17).

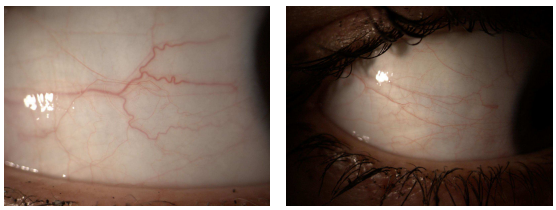


Figure 2.16: Example images where the camera is close to (left) and far from (right) the patient's eye in IMG dataset.

The images from IMG data set are larger than the ones from VID data set. The illumination and blurriness issues also appear in different areas. In the IMG set, the area of the superior eyelashes is slightly blurry and present some shadows created by the eyelashes. The VID set is affected near the left and right sides of the image, which often present slight to mild shadows. Bright spots caused by light sources appear

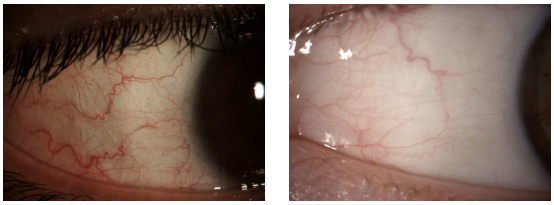


Figure 2.17: Example image in the VID data set (left) and IMG data set (right).

frequently on both data sets, and shadowed areas are less common in IMG than in VID.

The VID data set has been graded in both the Efron and the BHVI grading scales, while only the Efron scale evaluations are available for the IMG set. In general, the grades obtained in the IMG set are lower than the ones in the VID dataset.

It must be noted that the VID dataset was captured and evaluated in order to develop this methodology. The IMG dataset was used for a study regarding contact lenses comfort, and several parameters were measured, bulbar hyperemia among them.

To summarise, the absence of a standarised procedure for the image acquisition, even within the same dataset, causes a high variability. This variability alone is not enough to prevent the automatisation, but it will hinder the process considerably.

Chapter 3

Towards an automatic approach

Once the datasets had been established and analysed, and taking into account the drawbacks of the manual procedure, the automatic approach can be tackled by means of computer vision and machine learning algorithms through a series of steps.

The goal of this chapter is to serve as a summary of the automatisation process. First, the bibliographic study is presented, explaining the approaches that served as the basis for this work. Then, the automatic methodology that has been implemented is described and the main results of the work are discussed.

3.1 State of the art

To the best of our knowledge, there are no frameworks that propose a fully automatic approach to the problem at hand. Some semi-automatic approaches have been proposed [17, 18], which perform a manual selection of the region of interest, compute a few image features but do not describe the transformation from these values to the grading scale. In the work by Yoneda *et al.* [17] the main focus is the evaluation of the reliability and reproducibility of a bulbar hyperaemia grading software. In the study by Peterson and Wolffsohn [19], a comparison between an automatic measurement and a manual grading provided by experts is performed. The results highlight the sensitivity and reliability of automatic approaches versus subjective evaluation. However, the process is not fully automatic, as the steps for both the frame extraction and the conjunctiva

segmentation are not covered. Also, the data from objective and subjective approaches is compared, but the image features are not converted to any grading scale value.

Wu *et al.* [20] use images taken by a Keratograph as the input of the system. The goal is to compare a new corneal topographer to three subjective grading scales in order to assess its validity and reliability. They conclude that there is a significant correlation between the Keratograph's values and each of the scales. The Keratograph's algorithm returns the redness value in a 0-4 scale, with step = 0.1 and distinguishing between nasal and temporal. To that end, it uses a proprietary software, so the underlying process is unknown.

Tort *et al.* [21] and Wald *et al.* [22] both focused on the study of allergic conjunctivitis and its effect on hyperaemia level. They take into account only information about vessel morphology, such as width, density or tortuosity. They use each feature separately to obtain the final output of the system. They highlight that automatic approaches can find relevant parameters that are not evident for the human expert, improving the efficiency of clinical trials.

Moreover, in Downie *et al.* [23] a simple image feature, the percentage of red in RGB colour space, is compared with the evaluations of a group of optometrists and with the automatic evaluation provided by a Keratograph (R-scan). As this work is focused on how a simple objective measure can be compared with the manual approach, it does not include an automatic segmentation of the region of interest, nor the computation and combination of several features. A similar approach is taken in the work by Amparo *et al.* [24], where a framework is also proposed, but the human operator must manually adjust the level of white in the image and the region of interest. They propose also a computationally simple redness measure based on HSV colour space, and then remap it to the selected scale. The study shows that a correlation exists between the automatic approach and the grading scale.

Despite of the absence of a fully automatic methodology, several of the cited works obtain promising results when comparing automatic and manual evaluation, emphasising the importance of the automatic techniques to improve the clinical trials.

Additionally, there is a considerable number of more specific works, closely related to some of the individual steps involved in the process.

Regarding the first step for hyperaemia grading, obtaining the best frame of the video sequence, there are works in other fields that propose automatic methodologies for frame selection. In the study by Wolf [25], an algorithm for identifying key frames in video sequences using the local minima of the motion is proposed. Erol and Kossentini [26] also propose a key frame detection algorithm, but using shape information instead of motion. Even though these approaches provide interesting solutions for the frame selection problem, they show a strong dependence on the domain of the problem that they are solving.

In relation to the definition of a region of interest in the image, there are few works that tackle the automatic segmentation of the conjunctiva. One of these works is the one by Radu *et al.* [27], where an algorithm for sclera segmentation focused on biometrics is proposed and validated, obtaining good results. Unfortunately, the images that the authors use are vastly different to the images used for hyperaemia evaluation and, thus, this technique is not directly applicable.

There are also several articles on iris segmentation that propose different approaches that can be adapted to this environment. For example, in the study by Liu *et al.* [28], the iris boundaries are modelled as two circumferences by means of a Canny edge detector. Then, the authors locate the centre of the circle by computing the maximum value in the Hough space [29]. However, the images for iris segmentation are frontal or near-frontal views of the eye, and the eye is far from the camera, showing a more general view of the eyelids and other surrounding areas. Therefore, these proposals need to be adapted to side views of the eye. In the work by Kong and Zhang [30], the iris is defined as two circles, and the eyelids as two parabolas. The method is focused on the separation of the eyelashes and the reflections that appear within the image, as the authors comment how the shape assumptions help to get rid of the eyelids in most cases. In this domain, general assumptions about shape are not as effective as in the biometrics domain, since the images present more variability regarding shape and size. Still, there are several works about iris segmentation that can be taken as

starting points [31, 32, 33, 34, 35]. These approaches use a thresholding process as a base, and combine it with assumptions about the shape and location of the area they are segmenting. However, it must be noted that variable illumination and focus are expected to hinder the process. Also, the pupil and iris area present a different colouration to their surroundings, while an hyperaemic conjunctiva will present a similar hue to eyelid skin.

In order to evaluate hyperaemia, several parameters can be taken into account. Some examples are the general colouration of the conjunctiva, the number of visible blood vessels, and their widths [36]. Several works propose features that need to be calculated in order to evaluate the symptom. Papas [37] proposed several image features, including vessel quantity and hue characteristics. In this work, the correlation between the measures and the gradings was analysed, the highest value being obtained by a vessel quantity-related measure. Wolffsohn and Purslow [2] proposed features based on colour or edge detection. The validation was performed with the BHVI scale, and was more focused on the repeatability of the analysed characteristics than on performing the grading automatically. Park *et al.* [38] also proposed four image features related to red hue, vessel quantity and the area occupied by vessels. The authors validated their methods with two grading scales (consisting of 4 and 10 values, respectively). Results showed that one of the methods that measures the area occupied by blood vessels had the highest correlation with the expert gradings. The aim of these works was to either implement several image features or to analyse the relation of single features and experts' gradings. Thus, there is a lack of research regarding the comparison of features and the analysis of their interactions and their relevance in the literature.

There are also several works that depict the construction and validation of grading scales. It must be noted that there is an absence of any consensus about grading scale creation and how scales should reflect the experts' knowledge. In the work by Bailey *et al.* [39], the effects of scaling were tackled, and an explanation was given for how narrower scales provide more accurate results. The authors concluded that the specialists need training in order to use a wider range of values, but also that they are more confident when applying a smaller range of values. Several works have proposed

methods for scale construction and validation, such as [15], where a scale with 100 levels was put forward. The results showed that clinicians usually assign values as multiples of 5, effectively transforming the scale into one with less granularity. In the study by Fieguth and Simpson [14], the experiments were performed with another 100-step scale. The authors analysed gradings from 72 experts, and concluded that, although the grading is highly variable, it presented the same effects as those observed in the work by Schulze *et al.* [15], since selected values were mostly multiples of 5.

Finally, to the best of our knowledge, the transformation from the image features to a grading scale range is not tackled in other works. This is one of the most important and complex steps, as the transformation is far from straightforward. Yet, this step will provide the greatest insight into the experts' knowledge.

3.2 Objectives

Hyperaemia is an early symptom of several ocular pathologies. Some of these pathologies have a high incidence in the world population and, therefore, they have both high medical and economical repercussions. In this situation, the prompt diagnosis is a vital point in order to provide an immediate and effective treatment.

The main objective of this work is to develop a novel methodology that evaluates the hyperaemia level in the bulbar conjunctiva in a fully automatic manner. The inputs of the system are videos or images of the patient's eye. They show side views of the eye, with both sides of each eye represented. The methodology tackles four steps: the selection of the best frame of the video sequence (if necessary); the segmentation of the region of interest in the image, removing spurious information such as eyelashes or eyelids; the computation of the representative image features, such as vessel quantity or hue of the conjunctiva; and the transformation of these features to a grade in the scale. The output of the system is the value in the chosen grading scale. These main steps are represented in Figure 3.1.

All the stages of the hyperaemia evaluation process benefit from automation. The selection of the best frame of a video sequence reduces the invested time in a tedious task to a minimum. The segmentation of the region of interest ensures the objectivity

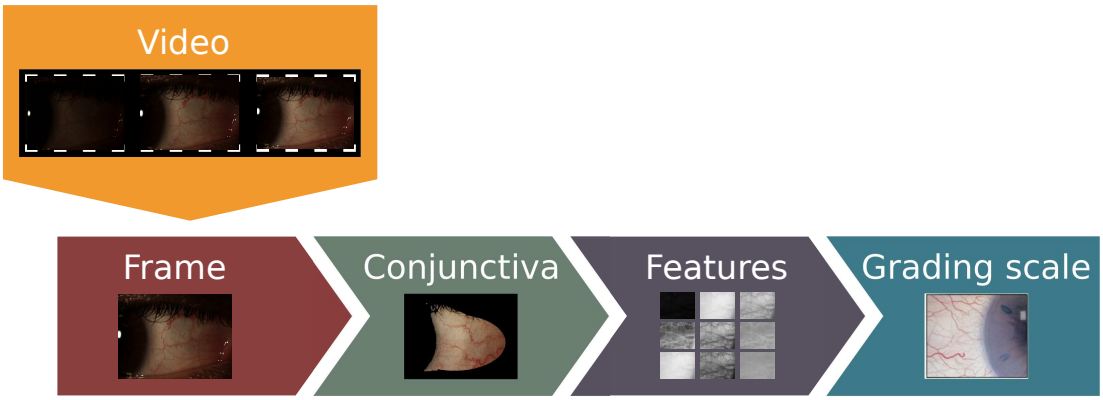


Figure 3.1: Steps for the automatic methodology: the input video is processed to select the best frame, the region of interest is segmented, several image features are computed and the values are transformed to a grading scale.

of the feature computation, as the used area is limited to the conjunctiva, and the same criteria is used to define this area in all the images. Moreover, the image characteristics cover a wide spectrum of possibilities, including knowledge from several experts and articles in the literature. Also, the combinations of characteristics are also studied, ensuring that the features included in the final system are relevant. Finally, the output provided by the machine learning techniques has the same range of values as that of the selected scale, which allows for a direct comparison with the experts' gradings. The objectivity and repeatability of the whole process is ensured, as the parameters of the methods and the underlying process are known.

Moreover, another objective of this work includes gaining a better insight in the experts' knowledge. This can be done by objectively identifying which characteristics are taken into account by the optometrists, and which ones among them are the most relevant for the grading.

3.3 Outline and main results

This work describes a fully automatic methodology for hyperaemia grading in the bulbar conjunctiva. The objective of the current section is to serve as a guide through the contents of the work. To that end, the steps that form the methodology as well as

the tests that were performed once it was completed are summarised, highlighting the main results and conclusions. For the sake of clarity, this section is divided according to the remaining parts of the work.

3.3.1 Grading hyperaemia

The second part of the work details the technical aspects of the automatic methodology as well as the experiments and results obtained during the development. It is structured in four chapters, one dedicated to each step of the proposed methodology.

Finding a suitable frame in a video sequence

If the input of the methodology is a video, the first step is to select the optimal frame of the sequence, that is, the frame that offers the best depiction of the bulbar conjunctiva. The chosen frame must show a clear view of the eye and have a good illumination, as Fig. 3.2 shows.

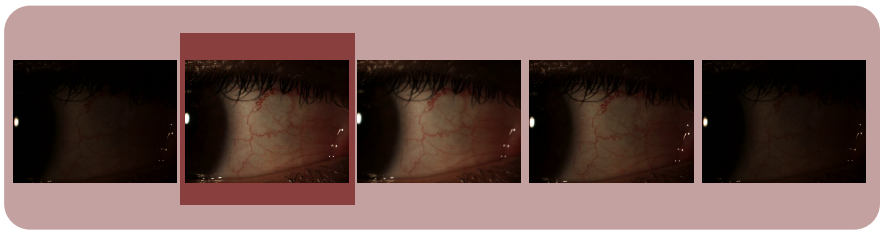


Figure 3.2: Input and objective of the first step of the methodology, the goal is to select the best frame of a video sequence.

To that end, the lightness of each frame is measured in order to select the frame with the best light conditions. Moreover, since the eye is not static through the video, unfocused frames are common. Thus, a blurriness measure is used to discard these unfocused frames. The proposed approach chooses a good frame for bulbar hyperaemia evaluation in 98% of the cases. Besides, in 90% of the cases, the selected frame is the best of the video sequence.

Chapter 4 presents the details of the implementation for this step of the automatic approach. On one hand, several illumination measures are proposed and applied to

this problem. On the other hand, blurriness metrics are defined in order to choose the best frame among the best illuminated frames.

Defining the region of analysis

Once the best frame is chosen, the second step is to limit the area of computation to the bulbar conjunctiva in a similar fashion as depicted in Fig. 3.3. Thus, it is necessary to separate the sclera and the vessels within from the eyelids and eyelashes. Both areas have different characteristics regarding hue, although high levels of hyperaemia can make the sclera appear similar to the surrounding skin. Moreover, as all the images are side views of the eye, some characteristics of the shape of the conjunctiva can be taken into account to facilitate the process.

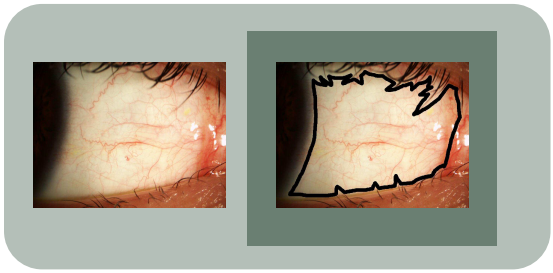


Figure 3.3: Input and objective of the second step of the methodology, the goal is to separate the bulbar conjunctiva from the surrounding areas.

Therefore, an exhaustive study of segmentation techniques has been conducted, including both state-of-art algorithms and ad-hoc approaches. Due to the variability of the images, the proposed approaches obtained different results depending on the characteristics of the data set where they were evaluated. Nevertheless, most of the segmentation approaches obtained acceptable results, with accuracy values above 0.7 in every case. In particular, the best techniques obtained an accuracy above 0.8 in both data sets. The eyelash area hinders the segmentation so an accuracy of 0.9 can be considered as the gold standard since it provides a good depiction of the central part of the conjunctiva. However, in order to increase the accuracy, several combinations of methods were proposed. By computing the outputs of ten segmentation algorithms and taking into account the points that are marked as conjunctiva by at least eight of

them, the sensitivity, specificity, precision and accuracy achieve values well above 0.8 in all the data sets. This approach, while more computationally costly, is much more stable and adaptable to new data sets. Furthermore, enhancement techniques were also investigated in order to improve the results. This way, the removal of white spots is recommended since the information of these points is lost and including them in the analysis can mask the real hyperaemia values.

Chapter 5 depicts the segmentation approaches that have been applied to this problem. This chapter presents the foundations of each algorithm and explains how they were adapted to the segmentation of the conjunctiva. Then, it presents the advantages of combining the outputs of several algorithms. Finally, the chapter analyses the influence of several enhancement techniques in the segmentation results.

Extracting information from the images

When grading hyperaemia, optometrists analyse the conjunctiva thoroughly. They search for several indicators of the symptom, such as thick vessels or a particularly red hue in the sclera. This expert knowledge can be modelled and applied by means of image processing techniques, as depicted in Fig. 3.4.

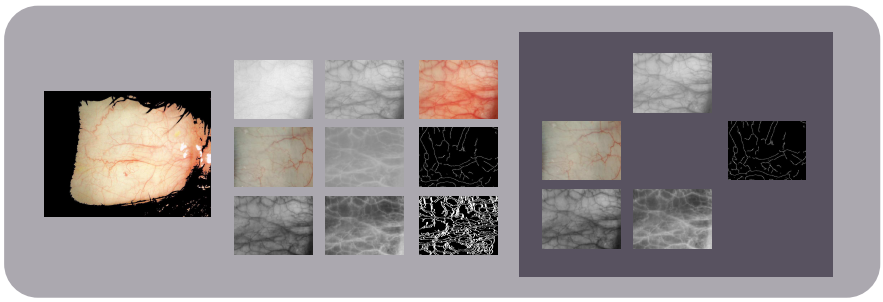


Figure 3.4: Input and objective of the third step of the methodology, the goal is to obtain several image features and to chose the best ones.

Thus, the third step of the automatic methodology starts once the conjunctiva region is defined. In this step, a series of image features is computed in that segmented region. A total of twenty-five features were computed, based on previous studies on the matter as well as information provided by optometrists. The features cover the vessel quantity or width and the hue in each part of the eye: the whole conjunctiva,

only the sclera and only the vessels. Moreover, as not all the features have the same significance depending on the side of the eye where they are computed, the twenty-five local features were also computed for both the nasal and the temporal conjunctiva, separately. Then, several feature selection techniques were applied, with the objective of reducing the feature set with minimal information loss. The most commonly chosen global features were those that take into account the level of red in the sclera or in the whole image in both datasets, although IMG dataset created larger feature subsets than VID and gave more importance to vessels. Regarding the feature selection with local and global features, the variability is higher, and the focus of IMG changes from vessels hue to background or vessel quantity, while VID shifts toward the vessel hue.

Chapter 6 defines the whole set of image features used in this work and analyses their relationship with the experts' evaluations. Moreover, several feature selection techniques are described in order to find an optimal subset of features that represents the hyperaemia levels. These steps are repeated to analyse the influence of local features and different datasets.

From the image features to the grading scale

Once the image features have been computed and an optimal set has been determined, several values are obtained for a given input. However, the relationship between these values and the grading scale levels is not straightforward. Once again, the difficulty to model the experts' knowledge hinders the process and, thus, machine learning algorithms are required to achieve the hyperaemia grade (Fig. 3.5).

Therefore, the fourth and final step of the methodology is to transform the optimal set of features to the values in the selected grading scale. As each grading scale has a different distribution and highlights different features, each one of them has to be tackled separately. First, the task was defined as a classification and a regression problem. To that end, techniques of both types were proposed and applied to the feature set, and their results, compared. The parameter that was used to evaluate the goodness of the algorithms was the mean squared error, that computes the differences between the experts' evaluations and the automatic outputs. The regression techniques

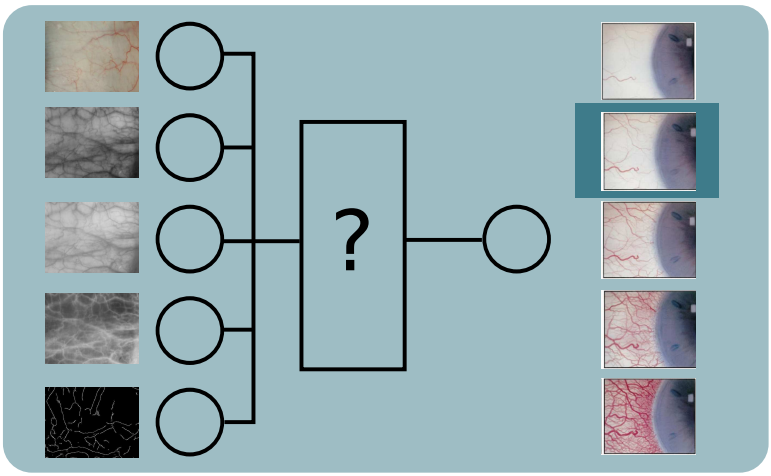


Figure 3.5: Input and objective of the last step of the methodology, the goal is to transform the image features in a value in the grading scale.

obtained the best results and, therefore, the regression approach was selected for further research. The VID dataset obtained the best results for the global features test with the whole feature set, a mean squared error of 0.048 and 0.041 in the Efron and BHVI scales, respectively. For the local and global features, the best results were obtained through feature selection, achieving a mean squared error of 0.058 and 0.046 in the Efron and BHVI scales, respectively. The IMG data set obtained the best mean squared error, 0.051, with feature selection in both global-only and global and local feature sets. As differences of 0.5 points in the scale between two clinicians are common, which is equivalent to a mean squared error of 0.25, the methodology is able to provide good results.

Chapter 7 describes briefly the regression and classification algorithms used during the last step of the methodology. Their results in the dataset are presented and discussed. This chapter also presents the results considering global and local features, as well as different datasets.

In order to summarise the contents of the second part, Figure 3.6 depicts the main results of this work for each step of the automatisation.

Figure 3.6: Main results obtained in each step of the automatic methodology.

3.3.2 Bringing the methodology to open scenarios

The third part of the work consists of three chapters that depict the behaviour of the methodology in real-world environments and how this methodology could be applied to new scenarios with, for example, different acquisition conditions or grading scales.

Repeatability of the methodology

One of the main issues that appear on medical imaging is the huge amount of variability that real-world images and videos present. As a result, even if a methodology or application works correctly during the development stage, it usually needs to be modified or tuned again to work in a different environment.

Therefore, a repeteability study was conducted with the objective of assessing the behaviour of each automatic step in the presence of certain alterations of the images, namely the presence of contact lenses or remains of blue dye. The results prove that the variability that the outputs of the automatic methodology experiment is similar to the variability of the human experts. Specifically, the average differences on evaluations of the same patient through consecutive checkups are 0.07 and 0.03 in the manual and automatic approaches, respectively.

Chapter 8 offers a complete repeatability study. To this end, both experts and automatic gradings are analysed in images of the same patient at different times and different conditions. For the automatic evaluation, the influence in each step of the methodology is also studied.

Class imbalance problems

When applying a methodology to a different environment, it is not uncommon that the available dataset is too small. Another regular occurrence is that not all the values within the range are well represented. Moreover, both situations can take place at the same time, which will have a negative impact on the results. In the context of bulbar hyperaemia, imbalanced data sets are common, as extreme values are a rare sight.

Thus, several data balancing techniques are studied in this work in order to minimise the problem. As the analysed dataset has few samples of each level of severity, alternatives that apply oversampling are preferred. The results show that these alternatives can improve the learning of the regression systems, reducing the MSE in more than 80%. Therefore, if the automatic methodology is applied in a different dataset that needs a re-training process, these balancing techniques should be applied. This way, the methodology could be useful in scenarios where some image prototypes are difficult to acquire.

Chapter 9 describes several data balancing approaches and how they can be applied to the hyperaemia imbalance problems. Particularly, this chapter is focused in how to split the continuous ranges into discrete classes in order to apply the class balancing techniques.

Precise segmentation

One major concern in image processing is the absence of standard procedures for the capturing process. The lack of standarised procedures has a negative effect on the development of automatic methodologies since the variability of the dataset is not delimited. In the context of bulbar hyperaemia, this implies that is not possible to ensure a universal segmentation approach that works with all the datasets. The position of the eye in the image or the distance to the camera can both change, and with them the best way to tackle the segmentation process.

Therefore, in this work the possibility of using a small central area that appears in all the images was explored. To that end, a central square of the image was used to compute the image features. Moreover, this square was subdivided, and the relevance

of each part was studied by means of feature selection and regression techniques. The results show that to use the central region, while having a loss of information associated in most cases, can still provide a MSE lower than 0.2, which is within the range of the experts' variability.

Chapter 10 depicts further details on the experiments to assess the relevance of a precise segmentation. This way, the chapter analyses the results of the methodology in a region of interest centred in the conjunctiva and even in smaller regions within this central area.

The main results for each of the studies on the application of the methodology to open scenarios are depicted in Figure 3.7.

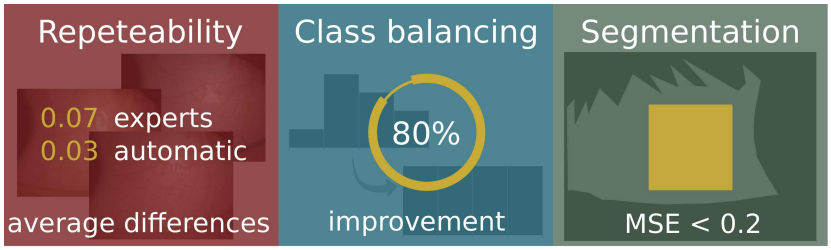


Figure 3.7: Main results obtained with the application of the methodology to open scenarios.

3.4 Further research

Despite the good results obtained by the proposed methodology, there are several lines of research that can be followed to improve and expand this work.

First, the publication of a large image dataset could be an interesting project. Currently, there are no public datasets of this kind of images. This hinders the development of assisted diagnosis tools, as well as the comparison of different techniques.

Next, the development of an automatic methodology for the analysis of hyperaemia evolution is another useful extension of the methodology. This analysis could be as simple as comparing the global output of the grading methodology through time or it could be more complex and focused on local changes. For example, a vessel could be highlighted in a picture but it could present a normal hue in the following checkup.

Thus, the analysis of local features must involve the computation of interest points in the image in order to apply registration techniques and align the local image features.

Finally, a user-oriented mobile application could be developed. A photo of the eye could be taken with the camera of a mobile device and then the methodology could be applied. This application could raise a warning if the picture shows a worrisome hyperaemia level, and then be used to easily monitor the changes over time. The benefits are specially remarkable in the risk groups of certain pathologies, as frequent checkups will allow them to detect the symptoms early. However, several topics need further study, such as the minimal requirements that the camera lens and lighting must fulfil in order to capture an adequate image of the conjunctiva. Also, it is necessary to analyse if common devices are powerful enough to perform the computations themselves, or if a server in the cloud is needed.

Part II

Grading hyperaemia

Chapter 4

Finding a suitable frame in a video sequence

One of the most common procedures for hyperaemia evaluation involves recording a video of the patient's eye. This approach presents the advantage of creating a large collection of images, so the most appropriate one can be selected to proceed with the grading. However, this decision is taken by following a time-consuming procedure, as the video must be analysed thoroughly in order to ensure that the selected frame offers a good depiction of the conjunctiva. A good frame must show as much conjunctiva as possible. Also, the illumination must be adequate, bright enough to observe the image features in detail. Moreover, as the eye will inevitably move during image capture, the specialist must check that the frame is not blurry.

This chapter is focused in the automation of the selection of the best frame of a video sequence by taking into account illumination and blurriness features.

The input of the system is a video of the patient's eye. The videos from the VID dataset start with a black frame, gain progressive illumination and then fade to black again. Moreover, there are occasional shadows caused by changes in the relative position of the eye in front of the camera, that can be misleading for the image processing algorithms. Therefore, to compute the lightness is advisable, as an adequate frame should be well illuminated. Also, the videos show both movement of the eye (blinking, looking around) and movement of the position of the eye regarding the camera. Ideally,

the selected frame should not show signs of blurriness derived of these movements, as blurriness will add noise in latter steps of the methodology.

4.1 Illumination

Table 4.1 shows the several implementations of lightness measures that were analysed: the RGB luminance, the V-channel from HSV colourspace, the L-channel from HSL colourspace and the L-channel from L*a*b* colourspace. Further details on the colour spaces that were used in this and the following chapters can be found in Appendix B. In RGB colourspace, there are three different formulations for the luminance, due to the differences in the perception that the human eye has and the implementation of the colour space [40]. The three formulations are: relative luminance/photometric approach for the colour spaces that follow the ITU-R BT.709 primaries, digital approach for the colour spaces that follow ITU-R BT.601 primaries and an alternative formulation of the latter. The ITU-R recommendations are a set of international technical standards developed in the International Telecommunication Unit (ITU). The BT.709 is for high-definition television, while the BT.601 is for standard-definition television. The lightness was calculated for each pixel in order to obtain the average value for the complete image.

Table 4.1: Lightness metrics.

Metric	Formula
RGB luminance	$Lum_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} 0.2126*R_{ij}+0.7152*G_{ij}+0.0722*B_{ij}}{n\times m}$
	$Lum_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} 0.299*R_{ij}+0.587*G_{ij}+0.114*B_{ij}}{n\times m}$
	$Lum_3 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \sqrt{0.299*R_{ij}^2+0.587*G_{ij}^2+0.114*B_{ij}^2}}{n\times m}$
HSV V-channel	$V_{hsv} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij}}{n\times m}$
HSL L-channel	$L_{hsl} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} L_{ij}}{n\times m}$
L*a*b L-channel	$L_{lab} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} l_{ij}}{n\times m}$

In the formulas, i, j represent the position of the current pixel. n, m are the rows and columns of the image, respectively. R, G and B are the channels for RGB colour space, V is the *value* channel in HSV colour space, L is the *lightness* channel in HSL colour space, and l is the *lightness* channel in L*a*b* colour space.

Since the objective is to chose the brightest frame, the frame with maximum lightness is selected as follows:

$$F = \underset{f}{arg\ max}\ \bar{L}(f) \tag{4.1}$$

where f is a frame and $\bar{L}$ is the average lightness measure computed in the whole frame. Figure 4.1 depicts the best frames obtained for each colour space for a given video.

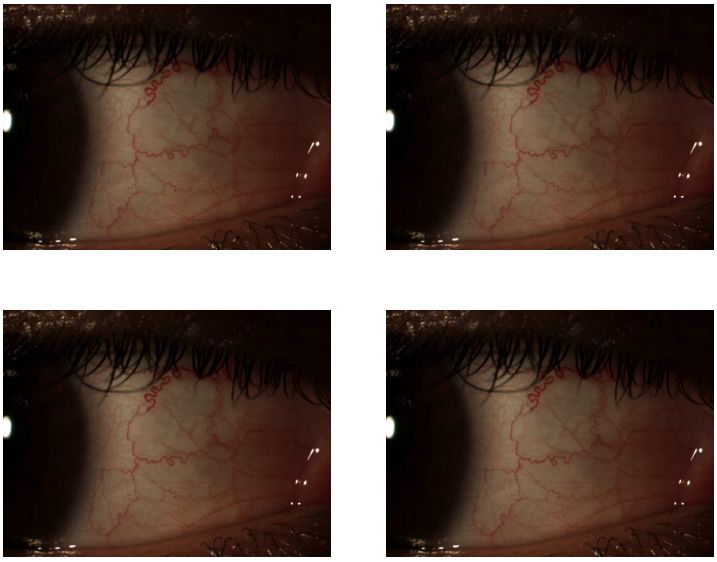


Figure 4.1: Selected frames using different colour spaces. From left to right and top to bottom: RGB, HSV, HSL and L*a*b*.

4.2 Blurriness measures

The selection of a blurry frame can lead to bias in several of the subsequent steps. Moreover, blurriness influences the hue of the area, and blends vessels and background, which adds noise to the image features. In order to correct this problem, the following blurriness measures were tested [41]:

Modified Laplace computes the sum of the absolute values of the convolution of an image with Laplacian operators.

Normalized variance calculates variations in grey level among the image pixels. It uses the power function, so it will emphasise larger differences from the mean intensity μ. Differences in average intensities among several images are compensated by dividing the variance by μ.

Tenenbaum gradient uses the Sobel operator to compute the image sharpness function.

Table 4.2 shows the formulas of the three blurriness approaches, where L_x and L_y are the Laplacian operators in each direction; I is the intensity of the image in each pixel; μ is the mean intensity in the whole image and S_x and S_y are the Sobel derivatives in each direction. Figure 4.2 depicts the the differences in the selected frame before and after taking into account the blurriness. The steps that conform this part of the automatic methodology are depicted in Fig. 4.3.

Table 4.2: Blurriness measures.

Metric	Formula
Modified Laplace	$B_{ML} = \sum_{i=1}^{n} \sum_{j=1}^{m} \lvert L_x(i,j) \rvert + \lvert L_y(i,j) \rvert$
Normalized variance	$B_{NV} = \frac{1}{mn\mu} \sum_{i=1}^{n} \sum_{j=1}^{m} (I(i,j) - \mu)^2$
Tenenbaum gradient	$B_{TG} = \sum_{i=1}^{n} \sum_{j=1}^{m} S_x(i,j)^2 + S_y(i,j)^2$

It must be noted that, in order to enhance the efficiency of the method, the blurriness is not computed for all the frames of the image. The system first computes the lightness and produces a subset of frames with the highest values. Then, the blurriness is computed in these frames. Moreover, there are some areas in the image that are not relevant for hyperaemia grading, such as eyelids and eyelashes. However, these areas affect the lightness or blurriness measures if remain a part of the computation. Therefore, a binary threshold is applied, prior to the blurriness calculation, in order to

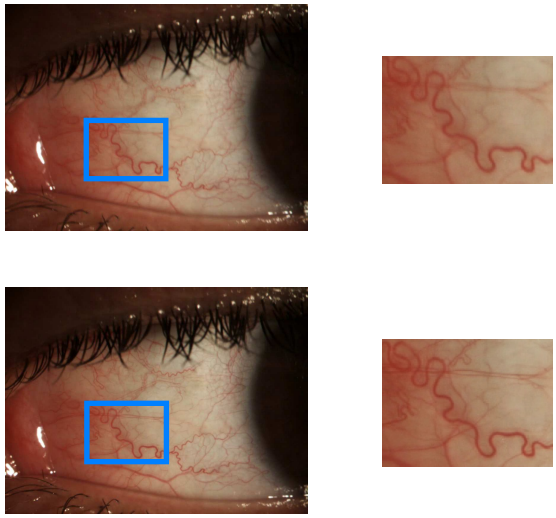


Figure 4.2: Detail of the blurriness of the image. Top: best frame without applying blur measures. Bottom: best frame taking into account image blurriness.

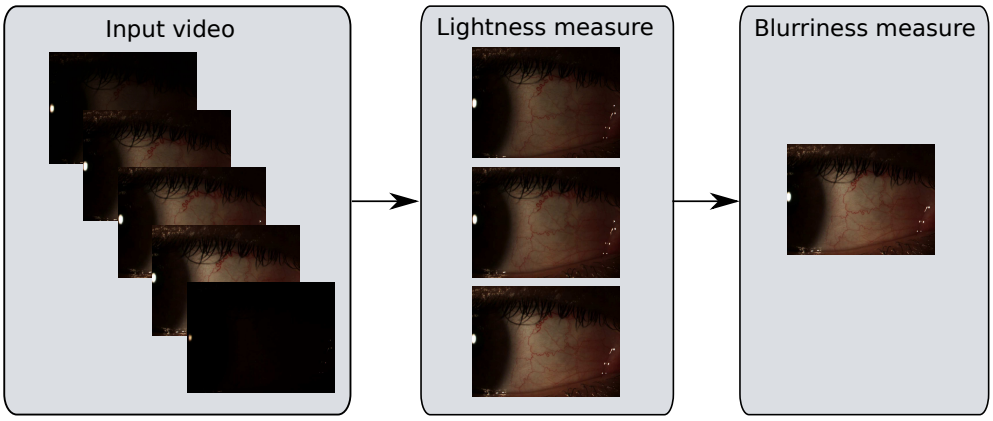


Figure 4.3: Steps conforming the frame selection, the first stage of the automatic methodology for bulbar hyperaemia grading.

remove them. This is a computationally efficient operation that restricts the measures to the conjunctiva and its surroundings.

4.3 Results

This methodology was tested in the VID_2 dataset. Regarding the lightness measure, the results were visually similar. Therefore, L*a*b* colour space was selected, since it offers the closest representation to the human vision [42].

In relation to the blurriness calculation measure, the Normalised Variance method is more than 10 times as fast as the other methods. However, as it is depicted in Figure 4.4, it presents problems in some cases. This was expected, as these methods measure the differences in intensity in the image. In the current environment, these differences are small and, therefore, insufficient to differentiate between sharp and blurry images correctly. However, Sobel and Laplace operators are focused on edge transitions. They assume that the changes in neighbouring pixels in a blurry image will be less drastic around the edges. Both methods provide good results in most of the videos, with a similar efficiency. The chosen algorithm was Tenenbaum gradient, which uses Sobel operator, because Laplace operator is more affected by noise than Sobel.

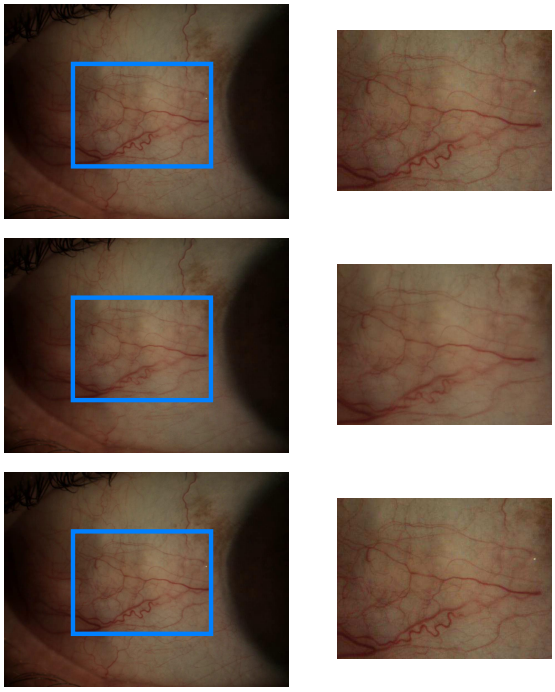


Figure 4.4: Detail of the best frame selected by the different blurriness measures applied after L_{lab}. Top row: B_{ML}, middle row: B_{NV}, bottom row: B_{TG}

Once the best lightness and blurriness measures were empirically chosen, they were applied to the 50 videos. The selected frames were presented to two specialists in order to validate the results. Each video and the selected frame was displayed to the optometrists. They were asked to answer two questions. First, if the selected frame was *optimal*, that is, if this is the best frame for the analysis. Second, in the cases where

the automatically selected frame was not the best, they were asked if it was *suitable*, this is, if, although there were better frames, the selected frame is accurate enough to evaluate hyperaemia. The results are depicted in Table 4.3. The column *Video issues* represents the videos that were reported as inadequate for hyperaemia evaluation by the specialist, due to poor quality.

Table 4.3: Validation of the frame extraction procedure.

Specialist	Optimal	Suitable	Non suitable	Video issues
E1	48	0	2	0
E2	43	4	0	3

In view of the data, the system obtains the best frame in more than 90% of the videos. Moreover, most of the discarded frames were also suitable for evaluation, although they were not the best ones of their video sequences.

Regarding efficiency, the time that takes to process a video of about 20 seconds long is 44.7 seconds on average in an Intel Core 2 Quad CPU (2.83 GHz) and 4 GB of RAM.

4.4 Conclusions

The first step of the automatic methodology for hyperaemia evaluation in the bulbar conjunctiva receives a video as input and selects the frame with the best illumination and the lowest blurriness. The methodology provides optimal results in more than 90% of the test set, and suitable frames in the 98%. Thus, it establishes an adequate setup for the subsequent steps of the methodology.

This procedure has the advantages of being objective and repeatable. Moreover, it is highly efficient, reducing most of the time that optometrists have to invest with the manual approach.

Chapter 5

Defining the region of analysis

The images and videos conforming the data sets depict the bulbar conjunctiva and its surroundings, this is, the eyelids, eyelashes, iris and pupil. All these regions represent spurious information that add nothing but noise to the system. Therefore, prior to the computation of image features, it is necessary to isolate the bulbar conjunctiva from the rest of the elements captured in the image.

This stage of the automatic methodology does not have a direct equivalence in the manual approach, and can be seen as a preparation of the image. However, it poses a special relevance within the process, as the inclusion of spurious information in the features can mask the real values. This is specially true when referring to the eyelids, as they can present a similar hue to the conjunctiva.

As the data sets are obtained from real world environments, illumination, blurriness or focus issues are fairly common, increasing the complexity of this step, which was high already, as the two areas being separated are similar in both hue and texture.

This chapter is focused on the segmentation of the region of interest in the bulbar conjunctiva. By observing the two grading scales that are employed in the data sets, the different focus of the prototypes becomes apparent. Moreover, specialists declare that they look at the whole conjunctiva when searching for distinctive features. Thus, the segmentation of the conjunctiva has to be as precise as possible, including most of the area.

First, both the state-of-art and ad hoc segmentation approaches are explained. Some of these approaches benefit from knowledge about the side of the image where the iris is located, as they will perform different operations on each vertical half of the image. Therefore, several automatic algorithms are detailed to solve this problem. In order to summarise the segmentation approaches explained in this chapter, their most relevant characteristics have been depicted in Fig. 5.1.

Further details on each approach can be found in this chapter. Also, several image enhancement methods are proposed, including colour constancy, which evens the illumination of the inputs; image filtering, that removes the noise; and bright spots removal, which removes the small bright reflection spots that are commonly found in the images. The final section of the chapter shows and discusses the results of each technique.

The VID_2 dataset was used as the basis for the development. Then, an additional validation with the IMG'_1 dataset was performed.

5.1 Segmentation of the bulbar conjunctiva

In order to separate the bulbar conjunctiva from other elements in the image, several approaches were studied, implemented and tested. They can be divided in three main groups, depending on the ideas they are based on:

Thresholding approaches use a thresholding as base operation. Several colourspaces were tested.

Shape-related approaches use information regarding the expected shape of the area that is being segmented in order to make assumptions and establish models.

Classic segmentation approaches include well known segmentation algorithms that do not fit in the other two groups.

5.1.1 Thresholding approaches

One of the most straightforward approaches to conjunctiva segmentation is performing a thresholding on the image. Thresholding approaches are fast and straightforward,

1

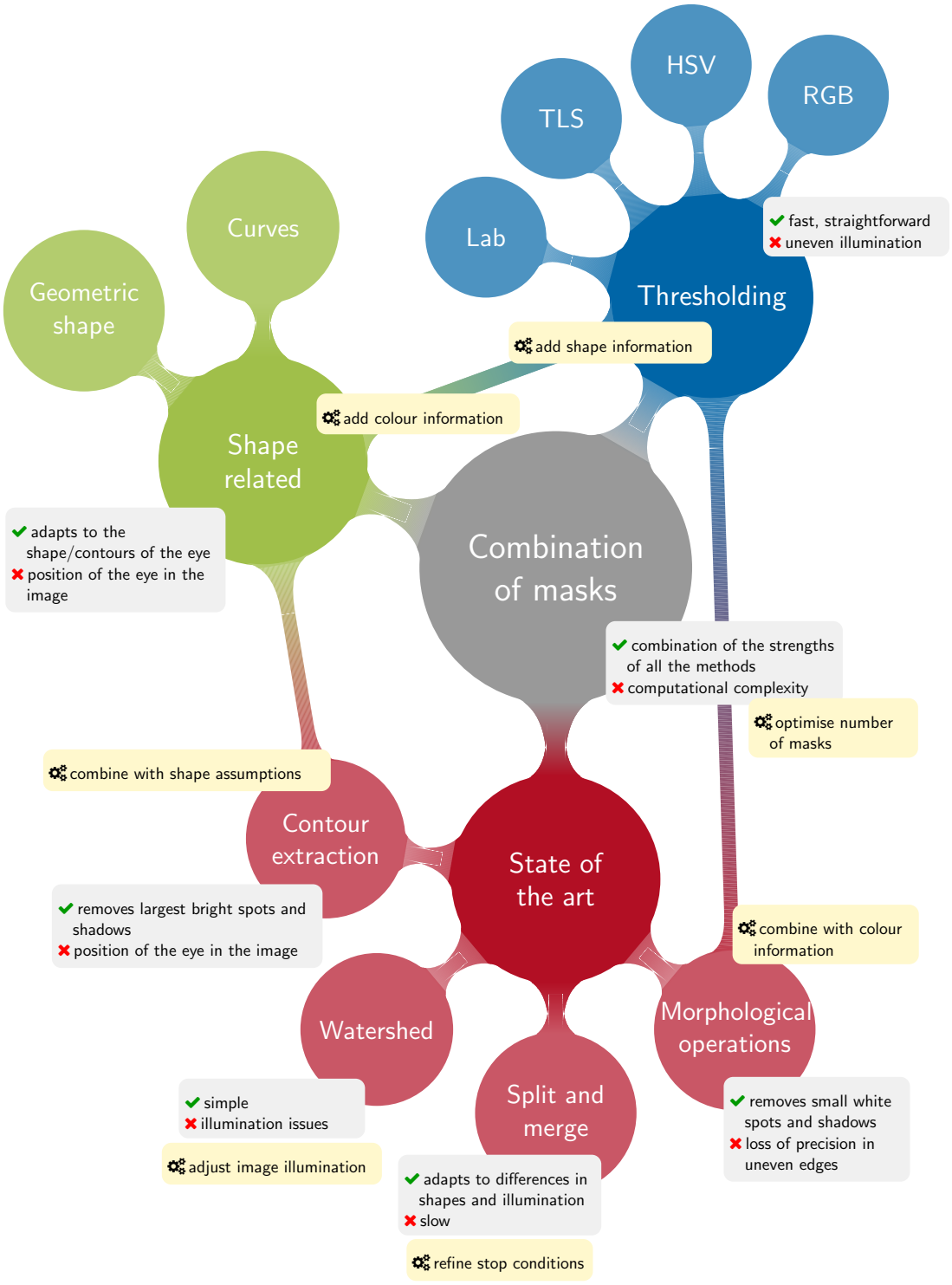


Figure 5.1: Main characteristics of the proposed conjunctiva segmentation approaches.

and can provide good results on segmentation scenarios [43]. Even in pictures that belong to non healthy individuals, the conjunctiva is expected to have lighter hues than the skin, pupil, and eyelashes. Several threshold values were tested but, due to the images' variability, the value that works best in most cases is the average intensity value. The following colour spaces were tested and its results are depicted in Fig. 5.2:

M_{TG}**.** Thresholding in the green channel of the RGB image using a fixed threshold value.

$M_{TG'}$**.** Thresholding in the green channel of the RGB image using the average of the intensity of the pixels in a certain part of the image as threshold. To that end, several grid divisions were tested as Fig. 5.3 shows. After analysing the results, the best outcome was achieved by using the mean intensity of the green channel in the central horizontal stripe of the image as threshold (configuration on the top left corner of the figure).

M_{TS}**.** Thresholding in the S channel of the TSL image using the average value in the channel as threshold. TSL colourspace [44, 45, 46] is commonly used to track skin areas in fields such as face recognition or gesture recognition. In this case, most of the unnecessary information contained in the images are the eyelid areas. However, as it is depicted in Fig. 5.2, hyperaemia makes conjunctiva colouration similar to skin colouration, lowering TSL's effectiveness.

$M_{TS'}$**.** Thresholding in the S channel of the TSL image using the average value as threshold. The level of red is also corrected to remove the vessel influence in the most severe hyperaemia images. To that end, an additional binary threshold is computed by taking into account the range of values that are close to red in HSV colourspace. Finally, an *or* operation is performed between the two masks. The second threshold is computed in HSV colourspace, and is obtained selecting only the values in the red range (H channel values 0-21 and 213-255, S and V values 96-255). Ideally, this mask retains the widest vessels, which can be removed by TSL thresholding.

M_{TV}. Thresholding in the V channel of the HSV colourspace. The V channel defines the brightness of the pixels so the mean value of the V channel is selected as threshold value. Therefore, the darkest parts of the image will be removed, as the conjunctiva region is typically the brightest area.

M_{TL}. Thresholding in the L channel of the L*a*b* image. In order to remove the areas closer to black, the threshold value is the average lightness value (L channel).

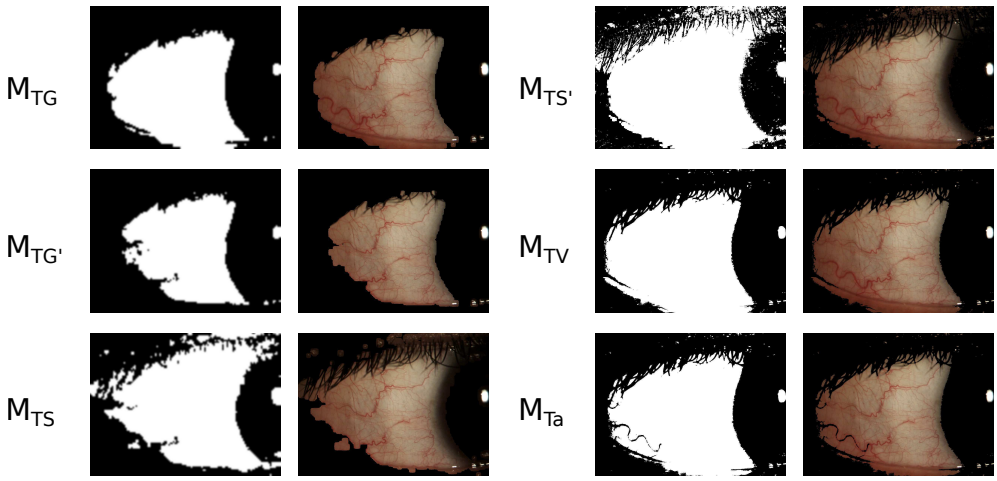


Figure 5.2: Application of the proposed thresholding approaches to the same image.

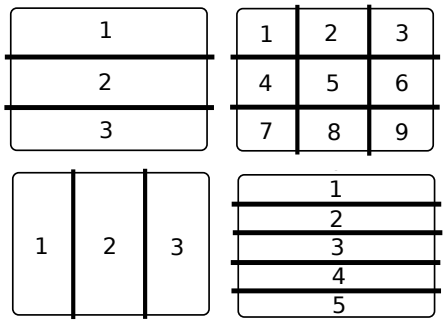


Figure 5.3: Different grids tested in the $M_{TG'}$ approach. After dividing the image in n fragments, one of them was chosen to compute the mean intensity, that will serve as a threshold for the complete image.

When the illumination was good, these approaches visually achieve good results. However, in uneven illumination conditions, a thresholding is not enough to segment

the conjunctiva, since only certain areas of the conjunctiva are segmented correctly, as Fig. 5.4 shows.

Figure 5.4: Results of the thresholding approach $M_{TG'}$ in an image with uneven illumination.

5.1.2 Shape-related approaches

Thresholding approaches are one of the simplest methods that can be applied to image segmentation. Their main advantage is their simplicity, as they are fast and provide results that are easy to understand. However, the characteristics of the images of the hyperaemia data sets are not optimal, and the edges of the conjunctival region commonly present issues that prevent the thresholding approaches from being fully effective. This way, the eyelids usually do not present a clear edge with the conjunctiva, and appear blended together instead.

One option is to rely on a preprocessing stage to improve the quality of the images. However, these enhancing techniques could not resolve all the issues. Another option is to provide additional shape information to the method, for example in the form of restrictions. By taking this approach, two types of segmentation are proposed: based on splines and based on ellipses.

Nevertheless, due to the differences in both vertical halves of the images, the shape-related assumptions need a previous step in order to be effective: to determine which side is the iris side. In the next section, several approaches to this problem are proposed.

Iris location

To distinguish the side where the iris is from the side where the corner of the eye/caruncle is can be highly beneficial for the segmentation. The shape of the region varies widely

from one side to the other, and differences on lightness are also apparent. In order to identify each side, a sensible course of action is to locate the iris region, as it has both a definite shape and hue. To that end, several methods were considered:

$IRIS_1$ This approach follows the idea that, taking into account the central part of the image, there should be a larger conjunctiva region in the iris side, due to the shape of the eye. First, an elliptical mask is created from the input in order to remove the corners of the image, as these areas are the most prone to present bright spots that could alter the results. The ellipse fits the image, so their major and minor axis are parallel to the x- and y-axis, respectively. A TSL thresholding (M_{TS}) was also performed in the input image and both masks are combined with a logical *and* operation. Then, the resulting image is divided in two vertical halves. The half with the largest number of white pixels is the iris side due to the eye shape. Figure 5.5 shows the steps involved in the $IRIS_1$ approach.

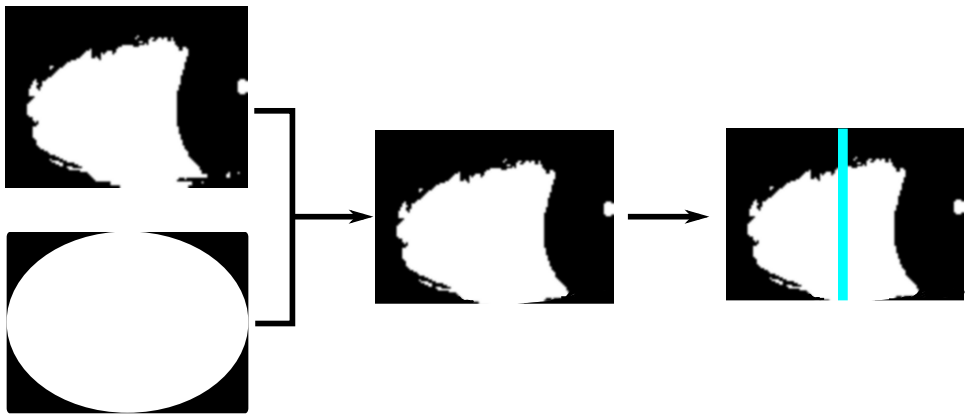


Figure 5.5: Steps conforming the $IRIS_1$ approach for iris location.

$IRIS_2$ This approach takes into account that the iris edges describe a smooth line, in contrast to the more complex pattern described by the corner of the eye. First, a threshold was performed in the green channel of the image. Then, a contour extraction algorithm is applied in order to find the largest contour in the image, that is a rough depiction of the conjunctiva edges. For both left and right borders of the image, the distance from the closest border of the image to the contour is stored in a vector. Then, the mean variance for both sets of candidates is

calculated, and their values, compared. The iris should present smaller variance, as the iris side presents a smoother curve than the corner of the eye. Figure 5.6 depicts the steps that conform the $IRIS_2$ approach.

Figure 5.6: Steps conforming the $IRIS_2$ approach for iris location.

$IRIS_3$ This approach assumes that the iris area describes an edge that can be modelled after a smooth parabola, while the corner of the eye does not. First, the corners of the image are removed with a mask as in the $IRIS_1$ approach. Then, a thresholding is applied in the green channel of the RGB image (M_{TG}) and the distances from the contours to the right and left borders are computed like in the $IRIS_2$ approach. Then, a parabola is fitted to both vectors, but only the function in the iris side would be close to a parabola. On one hand, if the caruncle/corner of the eye side is absent, this side will be represented as a straight line instead of a curve. On the other hand, if the corner of the eye is present, the curve on that side would be more acute and less smooth than the curve in the iris. Figure 5.7 shows the stages involved in the $IRIS_3$ approach.

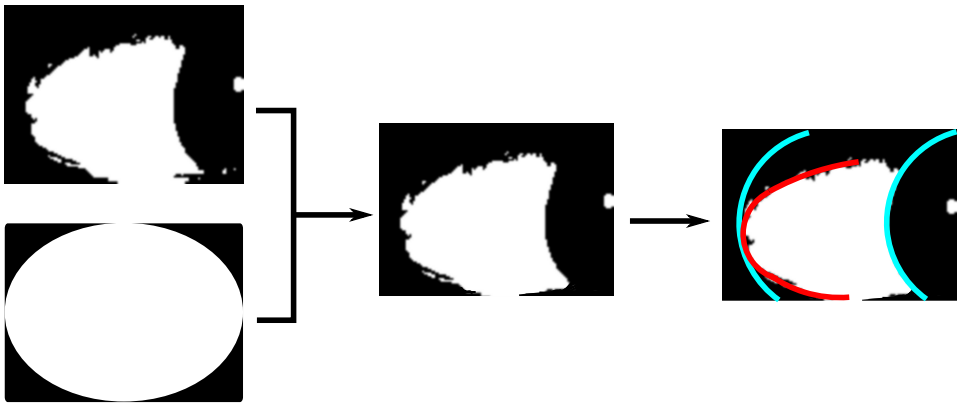


Figure 5.7: Steps conforming the $IRIS_3$ approach for iris location.

Spline modelling

Although the edges of the conjunctiva experiment a high level of variance, their underlying shape is similar in all the images. Therefore, each of the eyelids, as well as the iris edge, can be modelled after a smooth function, such as a spline. Thus, this approach will compute a series of reference points and use them to create the curves.

The first step is to locate the iris in the image as described in Section 5.1.2. Once the orientation is known, the next steps are focused on obtaining a simple contour of the conjunctiva and, finally, adjust it to the real edges.

For the second step, a binary threshold in the green channel of the RGB image is computed in the same manner than approach M_{TG}. Some reference points are computed depending on the side of the image: the extremes and centre of the iris, and the centre of the caruncle/corner of the eye. To that end, two vectors are created, one associated to the left side of the image and other to the right side, V_{left} and V_{right} respectively. These vectors store the distance from the image border to the closest contour point in each side of the image as shown in Fig. 5.8 (left). By combining this information with the orientation of the image, the four aforementioned reference points are found. For example, if the iris area is on the right side of the image, V_{right} is analysed in order to find two points (e_t, e_b) that verify the following conditions: the distance from e_t to the border of the image, and from e_b to the border of the image, must be lower than the mean distance of all the points in V_{right}, $\overline{V_{right}}$, and each one of them must represent a row located in the top and the bottom of the image:

$$\begin{cases} (e_{t_x}, e_{t_y}) = (V_{right}(i), i) \wedge i \in (1, \frac{length(V_{right})}{2}) \wedge V_{right}(i) < \overline{V_{right}} \\ (e_{b_x}, e_{b_y}) = (V_{right}(i), i) \wedge i \in (\frac{length(V_{right})}{2}, length(V_{right}) - 1) \wedge V_{right}(i) < \overline{V_{right}} \end{cases} \tag{5.1}$$

where (e_{t_x}, e_{t_y}) is the extreme obtained in the top rows of the image (upper eyelid) and (e_{b_x}, e_{b_y}), the one obtained in the bottom rows of the image (lower eyelid).

Once both extremes have been located, the centre is set as the middle point p between the two extremes. There is an additional restriction to p: its distance to the border of the image should be higher than the mean distance to the borders:

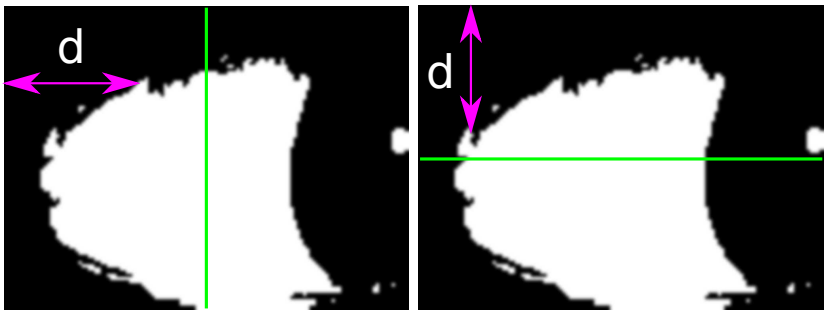


Figure 5.8: Distances to the closest border in the spline-based segmentation approaches. Left: horizontal distances to the closest vertical border. Right: vertical distances to the closest horizontal border.

$$(p_x, p_y) = (V_{right}(i), i) \wedge i = \frac{e_{t_x} + e_{b_x}}{2} \wedge V_{right}(i) > \overline{V_{right}} \tag{5.2}$$

If this restriction is not fulfilled, the closest point in the vector V_{right} that meets the criteria is chosen instead.

Regarding the opposite side of the image, the corner of the eye (c_x, c_y) is selected by following a similar procedure. The lowest value in the given distance vector (V_{left}) is chosen, as it defines the nearest white point to the image border. Two more reference points are needed, l_t and l_b, one in each eyelid. Labelling the image size as $m \times n$, a vertical line from $(m/2, 0)$ to $(m/2, n)$ is traced and the crossings with the segmentation mask are identified. There is a special situation that has to be considered: when the conjunctiva reaches the border of the image, there are no eyelids to take as reference. In this case, an additional vector V is computed. This vector, similar to V_{right} and V_{left}, also stores the distances for each image column, from the top (V_{top}) or bottom (V_{bottom}) border of the image to the closest white point. Once the distances are computed, the point is selected by searching the vector from the centre to the extremes, and selecting the first point with distance equal to zero (Fig. 5.8, right).

Finally, these six reference points (Fig. 5.9, left) can be used to draw three curves. First, a parabola is drawn for the iris, where the vertex is the iris centre and the extremes, the iris extremes. The two eyelids have a less clearly defined shape and, therefore, a second order polynomial is not accurate enough. Therefore, two sets of

n_p extra points are selected along each eyelid at a certain interval. These points are chosen by following the eyelid edge as depicted in Fig. 5.9, right.

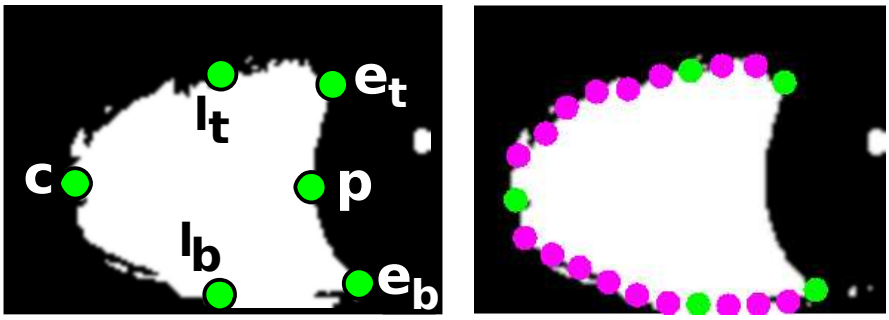


Figure 5.9: Reference points for the spline segmentation approaches. Left: extremes (e_t, e_b) and centre (p) of the iris region, corner of the eye/caruncle (c) and reference points in each eyelid (l_t, l_b). Right: extra points.

Once there are enough points, a spline is modelled for each eyelid. Splines are polynomial functions, defined piecewise on an interval. Each spline will start at the corner of the eye c_x and finish at the iris extreme e_b or e_t. Each spline S will be divided in k subintervals, where S will follow a given polynomial function P. Each pair of consecutive points in an eyelid will define a subinterval. The highest order of the set of P functions will determine the spline order. As cubic splines are used, each P and S will have order three. The obtained mask is depicted in Fig. 5.10 (top).

An alternative approach is proposed, with two major differences. Once the reference points have been computed, the contours of the image are obtained by means of the algorithm proposed in [47]. Since this algorithm can locate several connected regions, the smaller contours are removed. Moreover, the search of the remaining points for the eyelids is performed within a window of size w instead of taking into account only the neighbouring points. These changes handle the case where discontinuities prevented the location of auxiliary points, but in return made the result less adapted to the eyelids, as depicted in Fig. 5.10 (bottom).

This approach improves thresholding segmentations in poor illumination situations, as it is able to model the eyelids even when certain regions are missing. Spline approaches only need certain points to be present in order to create smooth curves. As

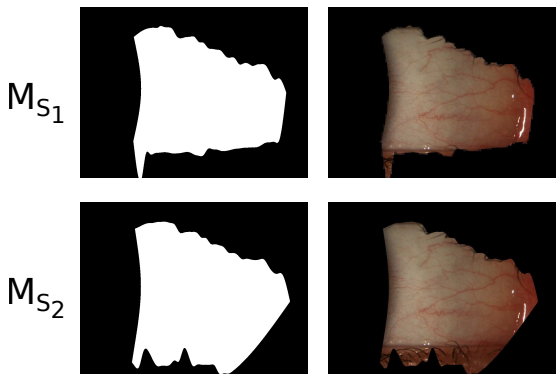


Figure 5.10: Application of the spline segmentation approaches to the same image.

the edges of the eyelids are generally smooth, the expected route that the spline is following is easy to predict. However, this approach does not solve the cases where illumination issues take place in the whole image. This is a direct consequence of using thresholding as the first step. If the initial image is not representative of the shape of the conjunctiva, the reference points can be misleading. Thus, a bias will be induced in the algorithms, producing suboptimal results. Moreover, it must be noted that splines are less accurate than thresholding when modelling the eyelashes. However, this is a minor complaint, as the area closer to the eyelashes is deemed as less relevant by the optometrists.

Figure 5.11 depicts the whole process for the spline-based approaches.

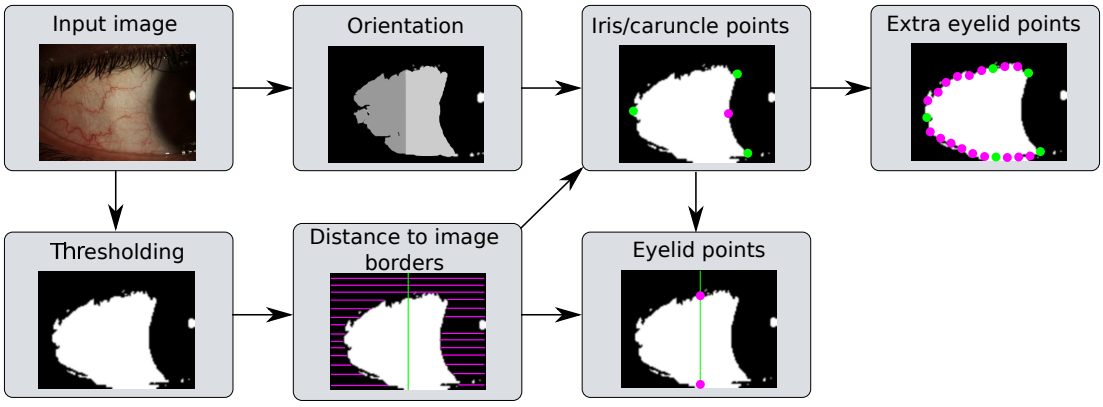


Figure 5.11: Steps conforming the spline based segmentation approaches.

Elliptical mask

A common approach in iris segmentation is to take into account the shape of the area that is being segmented, that is, the circular shape of the iris. This knowledge is used as a basis for the segmentation. This approach can be adapted to the particularities of the side view images of the bulbar conjunctiva. The closest simple shape that can be used to model the conjunctiva in this situation is an ellipse, except for the iris area, that will be tackled separately.

The first step to define an elliptical mask is to establish the location of its axes, the angle between them and the point where they cross. The equation of the ellipse is defined as follows:

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1 \tag{5.3}$$

where a is the radius of the major axis (x-axis), b is the radius of the minor axis (y-axis) and $(h,\, k)$ are the centre's coordinates.

In order to obtain an ellipse that comprises most of the conjunctiva, the major axis starts at the iris centre and ends in the corner of the eye. Regarding the minor axis, it starts and ends at the top and bottom eyelids respectively. Thus, the four reference points are obtained in the same fashion as that of the spline approaches. If one of the extremes of the major axis cannot be obtained, the middle point of the corresponding image border is selected instead:

$$\begin{cases} (e_{l_x}, e_{l_y}) = (\frac{n}{2}, 0) \\ (e_{r_x}, e_{r_y}) = (\frac{n}{2}, m) \end{cases} \tag{5.4}$$

for an image of size $m \times n$. Regarding the extremes of the minor axis, when a point that belongs to an eyelid is taken as extreme, the resulting ellipse was too restrictive. Thus, for a better representation, both extremes were shifted outwards a certain *margin* Δ_1, empirically determined. This assumption is depicted in the left image of Fig. 5.12.

The radius of the major and minor axis, a and b respectively, are then computed as follows:

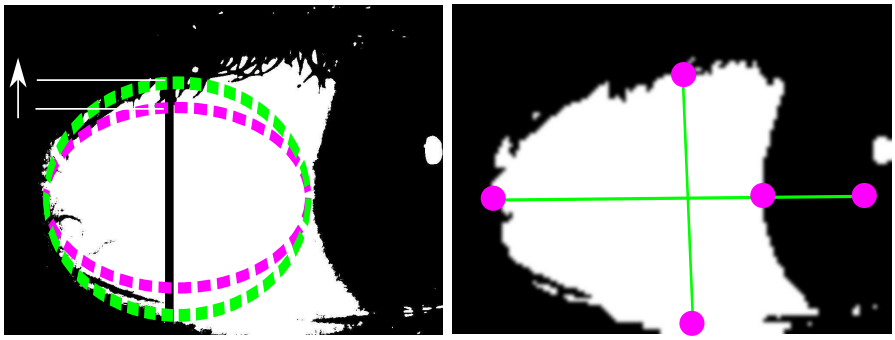


Figure 5.12: Shift points in the ellipse-based approaches in order to improve the modelling of the major axis.

$$\begin{cases} a = \sqrt{(e_{l_x} - e_{r_x})^2 + (e_{l_y} - e_{r_y})^2} \\ b = \sqrt{(e_{t_x} - e_{b_x})^2 + (e_{t_y} - e_{b_y})^2} \end{cases} \tag{5.5}$$

Finally, in order to ensure that the eyelashes within the conjunctiva are not taken into account, the ellipse is combined with a binary threshold t_{lashes} in the S channel of the TSL colourspace image by means of a logical *and* operation. As it is depicted in the top left image of Fig. 5.13, the obtained mask M_E offers an accurate representation of the corner of the eye/caruncle. Still, it has some undesirable effects, as the iris area is poorly represented because a large part of the closest conjunctiva is removed by the mask. In order to improve the representation of this side of the conjunctiva, the point associated to the centre of the iris was shifted in the major axis to the exterior of the eye:

$$\begin{cases} \begin{cases} (e_{l_x}, e_{l_y}) = (0, p_y) \\ (e_{r_x}, e_{r_y}) = (c_x, c_y) \end{cases} & \text{if } p_x \in [0, m/2) \\ \begin{cases} (e_{r_x}, e_{r_y}) = (n-1, p_y) \\ (e_{l_x}, e_{l_y}) = (c_x, c_y) \end{cases} & \text{if } p_x \in [m/2, m) \end{cases} \tag{5.6}$$

where (c_x, c_y) represents the corner of the eye and (p_x, p_y) the iris centre.

The top right image of Figure 5.13 shows the extended ellipse. It can be observed how the former issue is solved, as the area now includes the upper and lower regions

surrounding the iris. However, some fragments of the iris are also included within the mask, which will add noise in the next steps. Thus, the threshold t_{lashes} is used to improve the representation of the iris area. There are two possible approaches for the combination of the resulting mask of applying the threshold, M_{lashes}, and the shifted ellipse. One option is to apply a logical *and* operation, in the same manner as M_E. The resulting mask, M_{ED} is depicted in Fig. 5.14. The main drawback of this option is that the ellipse is too restrictive in the iris' surroundings, where the thresholding provides a better approach. Therefore, the second option is to perform a logical *and* operation only in the corner of the eye half of the image, and using the threshold output in the iris half. The obtained mask, M_{ET}, is depicted in the bottom left image of Fig. 5.13.

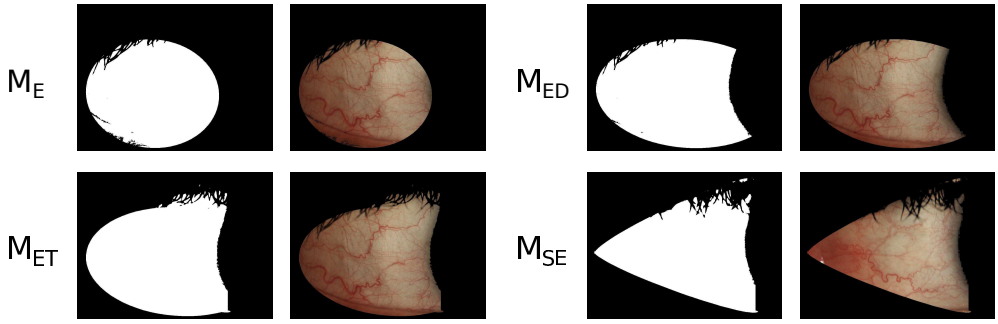


Figure 5.13: Application of the ellipse segmentation approaches to the same image.

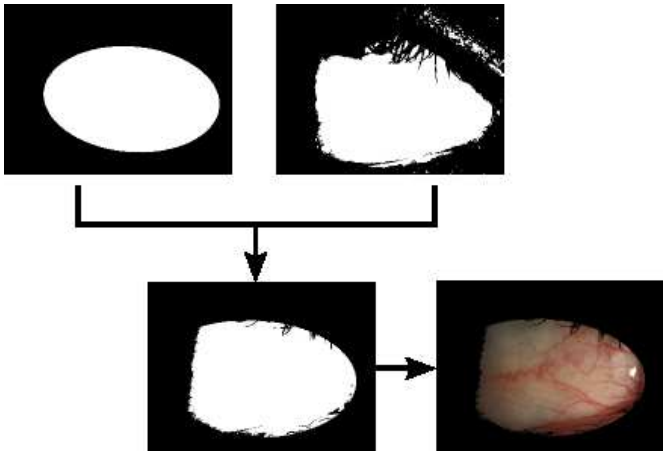


Figure 5.14: Segmentation of the bulbar conjunctiva by means of the combination of an elliptical mask and a binary threshold.

This approach presents several advantages over the aforementioned ones. Only four points are needed in order to determine the shape, which makes it less dependent on the accuracy of the previous thresholding. Also, the general shape and disposition of the images allows the use of general assumptions that obtain accurate results in most of the cases. However, when illumination issues are combined with location issues, neither the selection of the points nor the general assumptions are able to obtain acceptable results.

Lastly, a mask that models each eyelid as a fragment of an ellipse is created. To this end, two additional parameters are defined in each ellipse, starting and end angles. These parameters define which arc of the ellipse is drawn for each eyelid. For the major axis of each ellipse, it is necessary to locate the same two points: corner of the eye and iris centre. From the iris centre, the values used as extremes (e_t, e_b) when the iris is at the left or right side will be:

$$\begin{cases} (e_{b_x}, e_{b_y}) = (p_x - \Delta_2, n - \Delta_2) \\ (e_{t_x}, e_{t_y}) = (p_x - \Delta_2, \Delta_2) \end{cases} \quad \text{if } p_x \in [0, m/2)$$

$$\begin{cases} (e_{b_x}, e_{b_y}) = (p_x + \Delta_2, n - \Delta_2) \\ (e_{t_x}, e_{t_y}) = (p_x + \Delta_2, \Delta_2) \end{cases} \quad \text{if } p_x \in [m/2, m)$$

For the minor axis of each ellipse, the middle point of the x-axis is used to draw a line in a 90° angle to the nearest border (top or bottom) of the mask, and then to find the point in that line where the conjunctiva ends, i. e., the first black point from the centre of the ellipse outwards (Fig. 5.15). The bottom right image of Fig. 5.13 depicts the mask obtained with this approach.

This method depends only on the accurate determination of iris borders and corner of the eye, and it improves ellipse approaches in the iris extremes area. However, it will perform a worse modelling of the curves caused by the eyelashes in comparison to splines, though it will usually include less eyelid border. Figure 5.16 shows the steps involved in each of the ellipse-based approaches.

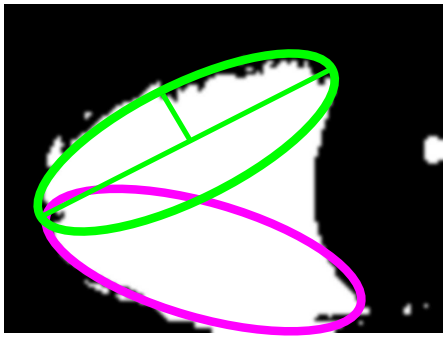


Figure 5.15: Determination of the minor axis of the ellipse that models the upper eyelid on the M_{SE} segmentation approach.

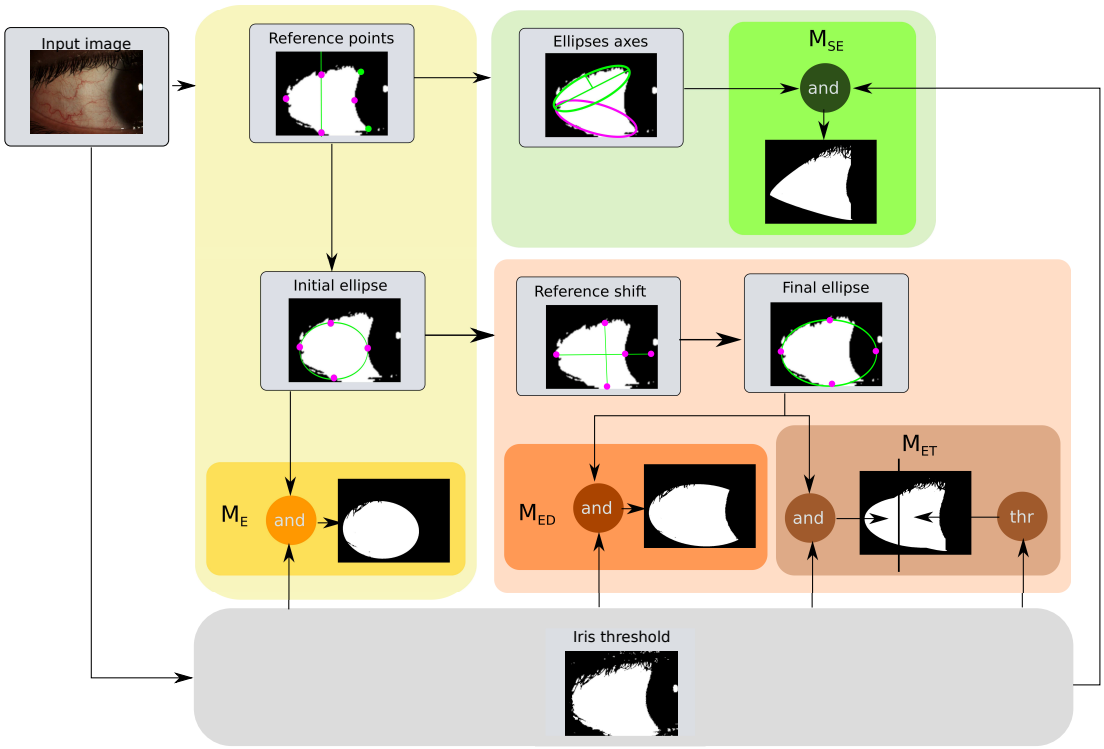


Figure 5.16: Steps conforming the ellipse based approach for conjunctiva segmentation.

5.1.3 Classic segmentation approaches

This section groups together other state of the art segmentation algorithms: morphological gradient, contour extraction, morphological opening, watershed, and split-and-merge. The first three approaches are based on simple operations, while the last two techniques are well-known algorithms, but also more computationally costly. Each

of the methods benefit from a rough threshold on the green channel of the RGB colourspace at some point, although they are not sensitive to the same issues that pure-thresholding approaches.

Morphological gradient

Morphological operations are another common technique that has been applied to edge detection [48]. A threshold t is applied to the input image in the green channel of RGB colour space (M_{TG}), and then the morphological gradient (difference between *erosion* and *dilation* operations) is computed for that thresholded image:

$$\begin{cases} X \ominus S = \underset{b \in S}{min}[f(x+b) - s(b)] \, \forall b \in S, x+b \in X \\ X \oplus S = \underset{b \in S}{max}[f(x-b) + s(b)] \, \forall b \in S, x-b \in X \end{cases} \tag{5.7}$$

where $f : X \rightarrow E$ is a grayscale image (X and E are the domain of the grayscale image and the range of gray values, respectively), $s : S \rightarrow E$ a grayscale structuring element (S is the domain of grayscale structuring elements), min is the minimum, max is the maximum and b is the structuring element [49].

Then, a contour extraction algorithm is applied [47] and the smallest ones are discarded. Several dilation operations are applied to the result in order to remove discontinuities. Next, the remaining contour is filled. Finally, the spurious regions created after the dilations are removed by means of a threshold. The obtained segmentation is depicted in Fig. 5.17 and Figure 5.18 depicts the steps of the methodology.

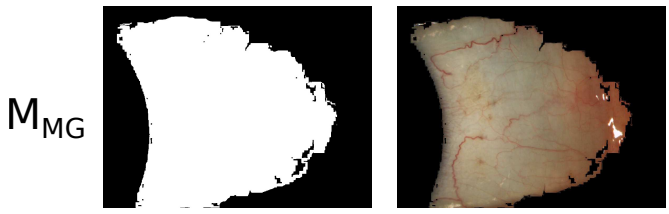


Figure 5.17: Segmentation of the bulbar conjunctiva by means of the morphological gradient approach.

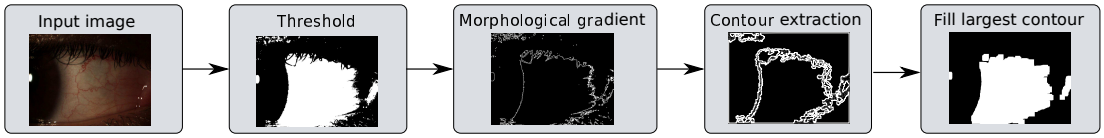


Figure 5.18: Steps conforming the morphological operation approach for conjunctiva segmentation.

Conjunctiva contours

The most common issue that thresholding approaches present is their low tolerance to bright regions. As these regions present a white hue, they are usually included in the segmented region. However, they present a common property, as they are smaller than the conjunctiva region. If they were as big as the conjunctiva, most of the image would present loss of information and, thus, would not be adequate for hyperaemia assessment. By taking into account the size of the regions, a new approach can be proposed by extracting the contours of the shapes that appear in a binary thresholded image and removing the smallest ones.

Therefore, the first step of the approach is to perform a thresholding in the green channel of the RGB image (M_{TG}). Then, the contours of the shapes are extracted by means of the algorithm proposed in [47]. The contours are stored as collections of vertices, which allows to compare their sizes. The smallest regions, which represent the bright regions, are removed. Then, the ones with the largest contours are filled to create the mask (Fig. 5.19).

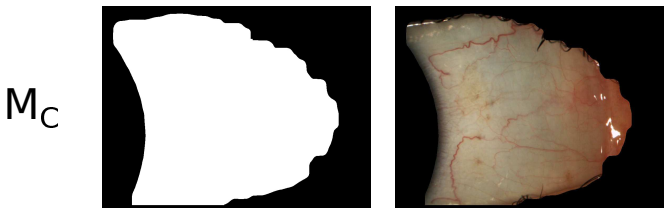


Figure 5.19: Segmentation of the bulbar conjunctiva by means of the contour extraction approach.

The benefits of this approach are specially remarkable for the images that present the largest bright points, specially when they appear in the edges of the eyelids, as these are the cases that include most noisy regions.

Morphological opening

This approach is based on performing series of morphological openings from an image thresholded in the green channel of the RGB colour space (M_{TG}). This operation removes the noise of the mask, i. e., those small regions that are kept by the threshold but do not belong to the conjunctiva (Fig. 5.20).

Figure 5.20: Segmentation of the bulbar conjunctiva by means of the morphological opening approach.

This approach present similar benefits than M_{MG}, and improves the thresholding based segmentations by removing the smallest silhouettes of the image.

Watershed segmentation

Watershed algorithms [50, 51, 52] depict the idea of a drop of water that flows following the gradient of an image, eventually reaching a local minimum. This principle has been used frequently to perform image segmentation. First, the image is transformed to grayscale, and next, a binary thresholding is performed. Then, a distance transform is applied, which labels each pixel taking into account the distance to the nearest boundary pixel. With this representation, the peaks of the image, which will serve as seeds for the watershed algorithm, are obtained. Then, the segmented areas are joined together, as some of the boundaries found after applying the distance transform correspond to blood vessels. Obtained results are depicted in Fig. 5.21.

In view of the segmented images, the watershed approach has some significant drawbacks. The first one is the location of the seeds, because if the conjunctiva has large, dark vessels, they will be marked as boundaries, splitting the result in many smaller regions. Then, when the algorithm is applied, the flood ignores the desired boundaries, and takes into account those defined by vessels or eyelashes.

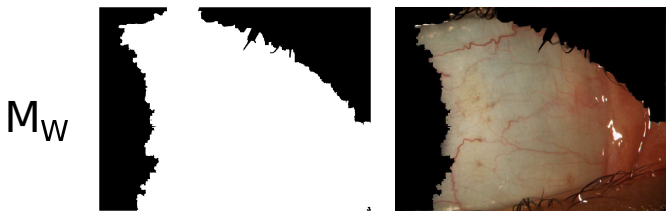


Figure 5.21: Segmentation of the bulbar conjunctiva by means of the watershed segmentation approach.

Split and merge segmentation

The split and merge segmentation [53, 54, 55] is based on a quadtree partition of the image. This is a tree-type data structure where each parent node has exactly four children. It is commonly used to divide in quadrants a two dimensional space. As a previous step to the application of the algorithm, a thresholding in the green channel of the RGB image, t_{SM}, is performed. The method needs also the definition of a stop condition h, which marks the end of the split part of the algorithm. For this implementation, said parameter is the standard deviation σ of the intensity of the image, with a threshold minimum value of t_h. Finally, in order to prevent the creation of smaller quadrants than necessary, a minimum block area a_m is also defined. The obtained mask and an example of application is depicted in Fig. 5.22.

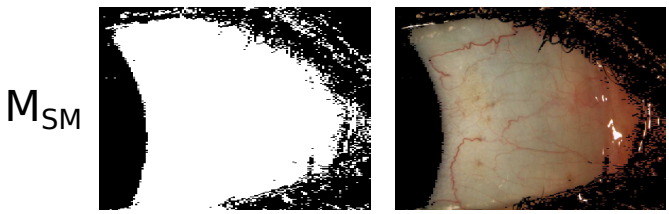


Figure 5.22: Segmentation of the bulbar conjunctiva by means of the split and merge approach.

This option provides a reasonable segmentation result, but has the drawback of being slow. For producing a good representation, it must repeat the process until fulfilling a strict homogeneity measure, so it will perform many split and merge operations.

5.1.4 Combination of masks

Due to the variability of the input images, there is not a single algorithm that performs well in all the cases. Some methods, such as the ellipse-based approaches, highly depend on the focus and the distance from the camera to the eye. Moreover, light skin tones can hinder the conjunctiva segmentation.

These circumstances piled together are expected to cause a worsening of the results in most methods. Therefore, a new approach was proposed in order to combine the strengths of each individual algorithm. The full set of segmentations was computed for each image. Then, a threshold t_n was established so that the pixels that belong to the segmented region in at least t_n segmentation masks are considered part of the conjunctiva, whereas the remaining pixels are considered background.

While it is potentially possible to find the perfect mask in each case by combining all the methods, this can be a time-consuming approach. Therefore, additional tests were conducted in order to find a suitable subset of masks that achieves good enough results and reduces complexity.

5.2 Enhancement techniques

The inputs of an image segmentation technique should meet some general conditions, such as regular illumination or absence of blurriness. Unfortunately, it is unusual for real world images to fulfil these requirements. Therefore, some image enhancement techniques were analysed in order to further improve the results. Two of these techniques, filtering and colour constancy, aim to improve the quality of the image prior to the segmentation process. The last proposed algorithm focuses on finding the brightest points of the image in order to remove them from the final segmentation.

5.2.1 Filtering

Filtering is a common step in image processing systems, as smoothing the image removes or minimises the noise. The following state-of-art algorithms were tested:

Gaussian blur (F_G) [56]. The image is smoothed by means of a gaussian function. As the space is two-dimensional, there will be two gaussian functions defined by the parameters σ_x and σ_y, computed from a given kernel size k_{size}.

Median filter (F_M) [57]. Each pixel of the image is replaced with the median value of the pixels inside the defined neighbourhood of size k_{size}.

Bilateral filter (F_B) [58]. This technique smooths images but preserves edges. It combines the ideas of closeness (spacial proximity of two pixels) and similarity (two pixels have photometric similar values). The final filter is a combination of a shift-invariant domain filter and a range filter, weighted by two gaussians with parameters σ_{colour} and σ_{space}.

Wavelet filter (F_W) [59]. n iterations of the wavelet transform are computed using Haar wavelets.

Figure 5.23 depicts and example of the application of the different algorithms to the same image.

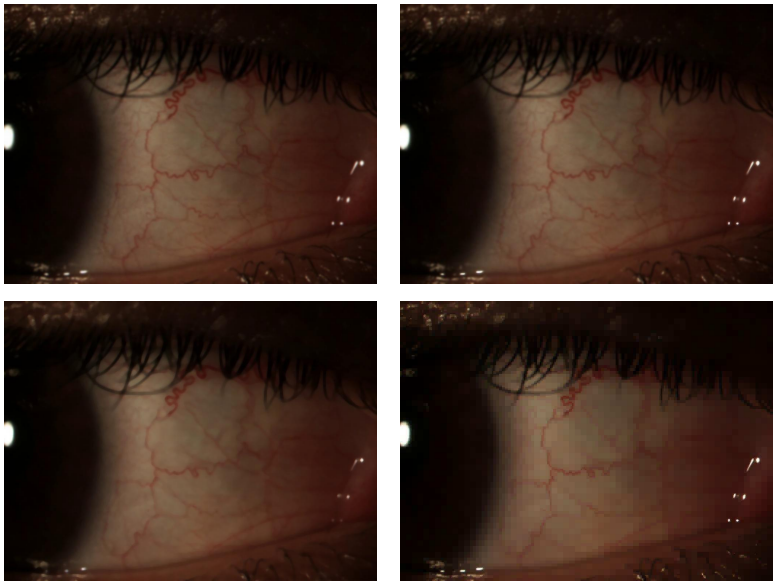


Figure 5.23: Effect of each filtering algorithm in the same image. From left to right and top to bottom: F_G, F_B, F_M, F_W.

5.2.2 Colour constancy

In the images conforming the data sets, each pixel is represented by three values. This property is known as trichromacity. Colour constancy algorithms [60, 61] are able to map these coordinates to a plane in order to make them illumination-independent. In this sense, the following algorithms were tested:

Grey world (CC_{GW}) [62, 63, 64]. The algorithm assumes that, under a white light source, the average colour in an image is achromatic. This implies that if the average colour changes from grey to other, it must be caused by the light source. Each channel of the image will be a combination of the real colour plus the illumination. The basic grey world normalisation can be defined as:

$$(\alpha R, \beta G, \gamma B) \rightarrow \frac{\alpha R}{\frac{\alpha}{p}\sum_i R}, \frac{\beta G}{\frac{\alpha}{p}\sum_i G}, \frac{\gamma B}{\frac{\alpha}{p}\sum_i B} \tag{5.8}$$

where α, β, γ represent the illumination variance and p, the number of image pixels. The algorithm computes the mean value for each channel i of the image I, $\overline{I}_i$, and the mean value for I, $\overline{I}$, as the mean of all $\overline{I}_i$. The applied transformation for each channel of the output image is:

$$CC_{GW_i}(x, y) = \frac{\overline{I}}{\overline{I}_i} I_i(x, y) \tag{5.9}$$

White patch (CC_{WP}) [65]. The algorithm assumes that the maximum response in an image (the maximum intensity) is caused by a perfect reflectance or white patch:

$$CC_{WP_i}(x, y) = \frac{255}{max(I_i)} I_i(x, y) \tag{5.10}$$

where I is the source image and i represents a concrete channel of the image.

White patch with a minimum threshold (CC_{WPt}). Values over the threshold t_{WP} are divided by the mean value of all the values over the threshold, and the values under t_{WP} remain unchanged:

$$CC_{WPt_i}(x, y) = \begin{cases} \frac{255}{\overline{I'_i}} I_i(x, y) & I_i(x, y) > t \\ I_i(x, y) & otherwise \end{cases} \tag{5.11}$$

where I'_i denotes the values of I_i that are higher than t.

Figure 5.24 depicts and example of the application of the different algorithms to the same image.

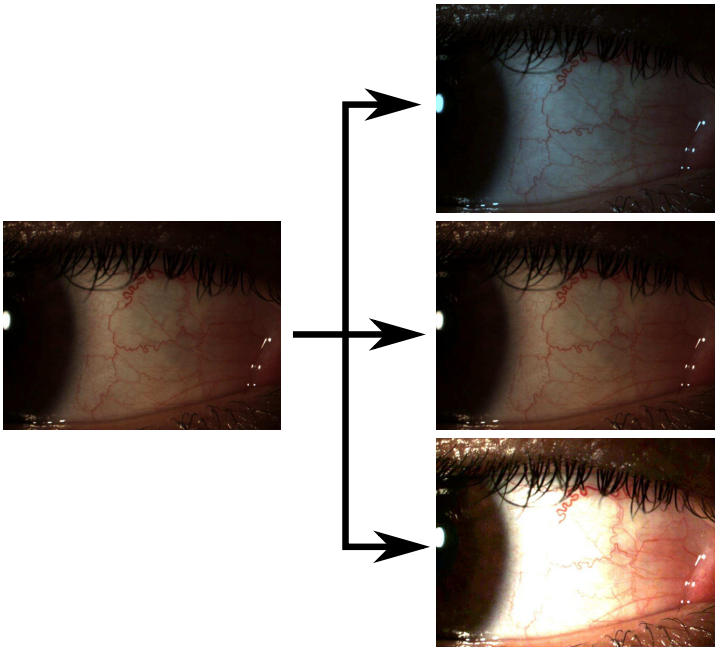


Figure 5.24: Effect of each colour constancy algorithm in the same image. From top to bottom: CC_{GW}, CC_{WP}, CC_{WPt}.

5.2.3 Shine removal

After observing the segmentation of the bulbar conjunctiva by different approaches, some common issues can be appreciated. The main drawback that most methods present is the occurrence of bright areas, caused by reflections of the light sources during the collection process. It can affect both the conjunctiva and its surrounding areas. The latter usually complicates the segmentation process for those methods based on hue or illumination. However, the former must also be tackled, as bright areas represent unknown information, and they must not be included when computing image features.

A property of these areas is that they are even brighter (and, thus, whiter) than the healthiest conjunctiva. Therefore, a binary threshold was applied to the green channel of the image, with an empiricall value t_{shine}. But this operation only obtains

the central part of the white areas, since the bright areas have a progressive change of colour instead of a clear edge. Thus, the area is enlarged by means of morphological operations. Specifically, an erosion operation is applied n_e times in order to ensure the full coverage of the area. Figure 5.25 shows examples of the shine removal application.

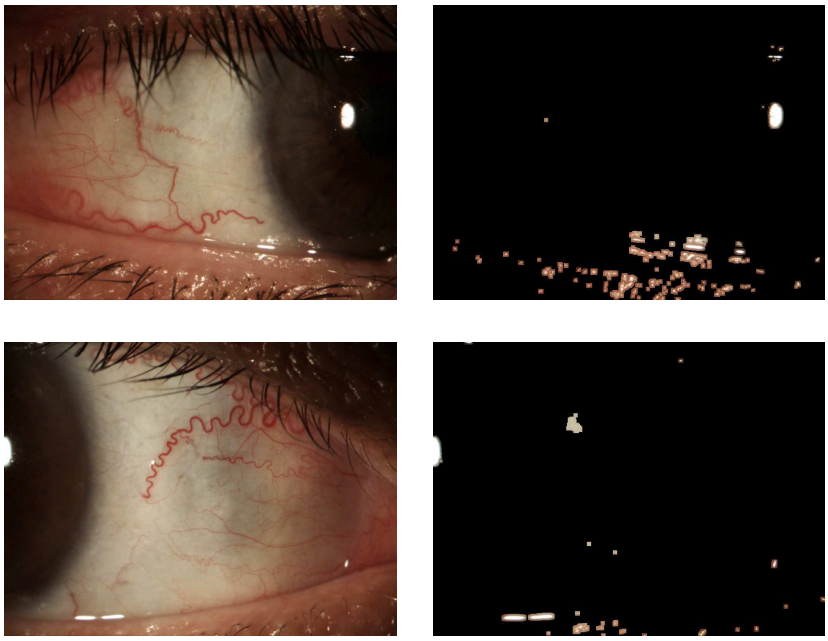


Figure 5.25: Application of the shine removal procedure.

5.3 Results

The following subsections depict the results obtained by each segmentation approach, as well as the impact that the enhancement techniques have in the results. First, the validation process that was followed to assess the performance of the algorithms is detailed. Then, the optimal parameters that were used for each method are listed. Finally, the results for each experiment are shown and explained.

5.3.1 Validation process

In order to establish a ground truth for the segmentation, an optometrist marked an image in order to depict which areas are observed during the grading. Figure 5.26 shows how the whole conjunctiva is taken into account.

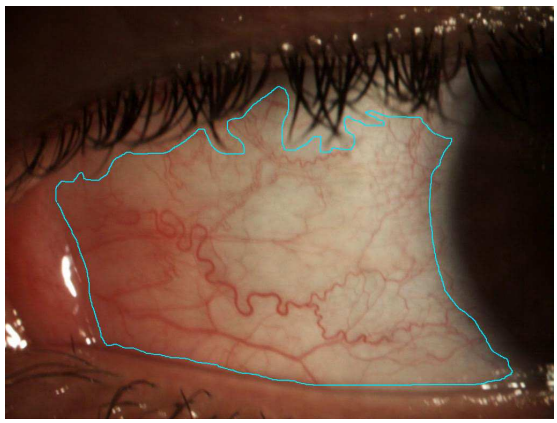


Figure 5.26: Area of the conjunctiva that specialists take into account when evaluating hyperaemia.

A subset of each image set was used to evaluate the conjunctiva segmentation procedure, in particular, VID_2 and IMG'_1. The former has been segmented twice, while there is only one manual segmentation for the latter. The first segmentation of VID_2 is smoother since it represents the edges of the lashes. The second segmentation is rougher and represents the shape of the conjunctiva with straighter lines. In order to validate the methods, the manual and automatic approaches are overlapped, and the following values are computed:

True positive (TP). A true positive is added for each pixel that belongs to the region of interest in both the manual and automatic masks.

True negative (TN). A true negative is added for each pixel that does not belong to the region of interest neither in the manual nor in the automatic mask.

False positive (FP). A false positive is added for each pixel that belongs to the region of interest in the automatic mask, but not in the manual one.

False negative (FN). A false negative is added for each pixel that belongs to the region of interest in the manual mask, but not in the automatic one.

By summing and averaging the values, these characteristics were computed for the whole image set. Then, these values are used to compute the following statistical measures, that allow the comparison of the methods:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.13}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.14}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.15}$$

It is important to remark that there are several equally valid manual segmentations for each image. This fact can affect the validation results, since to compare with one or another manual segmentation will change the values of the statistics. As the average accuracy of the comparison of both manual segmentations for VID_2 dataset is closer to 0.9 than 1.0, the gold standard for the segmentation process will be 0.9.

As the areas around the edges of the conjunctiva are deemed less important by the optometrists, a *conservative* method is more desirable, namely those methods that give priority to discard spurious regions than to include all the sclera. Thus, it is important to minimise the number of FP and, therefore, to obtain a high specificity value.

5.3.2 Parameters

Exhaustive tests were performed with the VID_2 dataset in order to find the best parameters for the implemented algorithms. A set of parameters were considered better than another if their accuracy was higher while no statistic dropped below 0.7. If there were no parameters that achieved at least 0.7 at a time in the four statistics, only the best accuracy was taken into account. Table 5.1 shows the best parameters found for each algorithm. In the equations, μ represents the average value of the given channel, Δ represents the margin or difference between the reference point found by the ellipse-based algorithms for the extremes of the axes and the shifted point, R represents a range of values, θ represents an angle, n represents a number of operations, k is a kernel size, t are the threshold values, a is the maximum area for the split stopping condition, σ stands for standard deviation and, finally, σ_n represents the parameters of each gaussian function in the bilateral filter.

Table 5.1: List of parameters used in the segmentation algorithms.

Type	Method	Parameters
Segmentation	M_{TG}	$t = 100$
	$M_{TG'}$	$t = \mu_{G'}$
	M_{TS}	$t = \mu_S$
	$M_{TS'}$	$t = \mu_S$, $R_{red} = (0, 96, 96)$ to $(21, 255, 255)$ and $(213, 96, 96)$ to $(255, 255, 255)$
	M_{TV}	$R_{black} = (0, 0, 0)$ to $(255, 255, t)$, $t_V = \mu_V$
	M_{TL}	$t = \mu_L$
	M_{S_1}	$t_S = \mu_G$, $n = 50$
	M_{S_2}	$t_S = \mu_G$, $n = 50$, $w_{size} = 7 \times 7$
	M_E	$\Delta_1 = 250$
	M_{ED}	$\Delta_1 = 250$
	M_{ET}	$\Delta_1 = 250$, $\theta_{start} = 90$, $\theta_{end} = 270$, $t_{TSL} = \mu_S$
	M_{SE}	$\Delta_2 = 20$
	M_{MG}	$t = 100$
	M_C	$t_C = 100$
	M_{MO}	$n_{iter} = 20$, $n_{oper} = 20$, $t = 100$
	M_W	$t = 40$
	M_{SM}	$t_{SM} = 40$, $t_h = 5.8$, $h = \sigma$, $a_m = 25$
Filtering	F_G	$k_{size} = 11 \times 11$
	F_M	$k_{size} = 11 \times 11$
	F_B	$\sigma_{colour} = 250$, $\sigma_{space} = 250$
	F_W	$n_W = 4$, $t_W = 50$
Colour constancy	CC_{GW}	-
	CC_{WP}	-
	CC_{WPt}	$t = 30$
Shine removal	S	$t_{shine} = 200$, $n_e = 5$

5.3.3 Identification of the image orientation

The results for the three proposed methods for computing the image orientation and both image sets are depicted in Table 5.2. The methodology does not take into account the eye or side that the image belongs to. However, for validation purposes, the iris on the right of the image (from the observer's point of view) is considered a positive and the iris on the left of the image, a negative. Thus, a TP is added if the method correctly establishes that the iris is on the right side, a TN when both method and ground truth set the iris on the left side, a FP when the method claims that the iris is on the right side but it is on the left side instead and, finally, a FN when the automatic approach marks the iris in the left side but it is located in the right side instead.

Table 5.2: Orientation computation results.

	VID dataset								IMG_1 dataset							
Method	TP	TN	FP	FN	Sens.	Spec.	Accu.	Prec.	TP	TN	FP	FN	Sens.	Spec.	Accu.	Prec.
$IRIS_1$	81	79	3	0	1	0.963	0.982	0.964	0	64	81	0	-	0.441	0.441	0
$IRIS_2$	28	56	26	53	0.346	0.683	0.515	0.519	0	58	87	0	-	0.400	0.400	0
$IRIS_3$	10	32	53	68	0.128	0.377	0.258	0.159	61	35	24	25	0.709	0.593	0.662	0.718

The images of the IMG'_1 image set frequently present a total absence of eyelids and hence, the results obtained with methods based in delimiting their contours are unsuitable. The iris is present in most images but the vertex area is usually merged with the border of the image.

The $IRIS_1$ method works almost flawlessly when the edges of eyelids or eyelashes appear in the image. If the eye is closer to the camera, and the surroundings of the conjunctiva are not depicted, the best performance is achieved with $IRIS_3$, although the results are suboptimal.

5.3.4 Segmentation of the bulbar conjunctiva

The results for the thresholding approaches in each segmentation of VID_2 dataset, as well as in the segmentation for IMG'_1 dataset, are depicted in Table 5.3. The differences

between datasets are clear, as it can be observed in the M_{TG} approach, that obtains a high specificity in VID_2 dataset with a low sensitivity, and the opposite values in the IMG'_1 scenario. Although some characteristics are more desirable than others, a minimum value for each measure was established in 0.75. Therefore, approaches M_{TV} and M_{TL} are the most suitable in both datasets.

Table 5.3: Sensitivity, specificity, accuracy, and precision for each threshold-based segmentation procedure.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
Mask	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_{TG}	0.682	0.863	0.750	0.880	0.740	0.851	0.790	0.850	0.896	0.652	0.798	0.811
$M_{TG'}$	0.710	0.894	0.777	0.889	0.781	0.882	0.820	0.858	0.618	0.978	0.746	0.975
M_{TS}	0.844	0.704	0.777	0.784	0.874	0.664	0.762	0.714	0.910	0.750	0.841	0.846
$M_{TS'}$	0.903	0.573	0.752	0.733	0.930	0.536	0.726	0.664	0.960	0.452	0.761	0.737
M_{TV}	0.801	0.830	0.801	0.850	0.870	0.813	0.829	0.812	0.777	0.848	0.788	0.878
M_{TL}	0.784	0.880	0.814	0.884	0.851	0.857	0.842	0.843	0.796	0.895	0.818	0.910

Regarding the spline approaches, the values obtained in both implementations are depicted in Table 5.4. The approach M_{S_1} is far superior to the other one, although it does not meet the criteria of achieving more than a 0.75 in all the parameters and, therefore, none of the segmentations were considered good enough.

Table 5.4: Sensitivity, specificity, accuracy, and precision for each spline-based segmentation procedure.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
Mask	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_{S_1}	0.722	0.887	0.789	0.871	0.790	0.870	0.828	0.832	0.795	0.761	0.771	0.826
M_{S_2}	0.498	0.936	0.675	0.885	0.551	0.927	0.725	0.855	0.002	0.947	0.364	0.075

The results of the four ellipse-based approaches are depicted in Table 5.5. There are high discrepancies between datasets in this case, as M_{ET} offers good results in VID_2 dataset, while the values for IMG'_1 dataset are poor in all the ellipse-based methods. The results show how the methods that need to make assumptions on the shape of the conjunctiva have a deficient performance on the IMG'_1 dataset. This is caused by the

absence of conjunctiva edges in this set, which hinders the stage of determination of reference points. Moreover, as the eye is closer to the camera, the conjunctival region does not resemble an elliptical shape anymore.

Table 5.5: Sensitivity, specificity, accuracy, and precision for each ellipse-based segmentation procedure.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
Mask	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_E	0.589	0.973	0.744	0.964	0.649	0.956	0.792	0.932	0.019	0.994	0.387	0.665
M_{ED}	0.700	0.950	0.795	0.945	0.766	0.928	0.833	0.908	0.018	0.995	0.387	0.718
M_{ET}	0.759	0.934	0.824	0.933	0.823	0.904	0.853	0.889	0.298	0.948	0.542	0.760
M_{SE}	0.680	0.962	0.794	0.956	0.738	0.934	0.829	0.912	0.389	0.814	0.554	0.629

The results for the segmentations obtained with the remaining methods are depicted in Table 5.6. Once more, discrepancies can be observed in some approaches. For example, M_C obtains a specificity above 0.8 in VID_2 dataset, while the same parameter is below 0.1 in IMG'_1 dataset.

Table 5.6: Sensitivity, specificity, accuracy, and precision for each uncategorised segmentation procedure.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
Mask	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_{MG}	0.764	0.912	0.819	0.910	0.846	0.900	0.865	0.886	0.810	0.903	0.830	0.918
M_C	0.837	0.842	0.829	0.869	0.906	0.815	0.854	0.827	0.999	0.096	0.653	0.642
M_{MO}	0.797	0.886	0.827	0.893	0.879	0.874	0.870	0.867	0.982	0.616	0.838	0.801
M_W	0.660	0.811	0.714	0.829	0.714	0.798	0.746	0.790	0.948	0.352	0.722	0.709
M_{SM}	0.818	0.829	0.814	0.854	0.877	0.800	0.832	0.805	0.797	0.907	0.829	0.924

Regarding the computation times, the fastest approaches are the thresholding in RGB colourspace (M_{TG} and $M_{TG'}$), that take about 0.005 seconds in an Intel Core 2 Quad CPU (2.83 GHz) and 4 GB of RAM. The other thresholding methods are slower, but take 0.26 seconds at most. The shape-related approaches, as well as M_{MG}, M_C and M_{MO} take from 1.3 to 1.6 seconds on average, except for M_{ET}, that takes 3.5 seconds. Finally, the slowest approach is M_{SM}, that takes 6.7 seconds on average.

5.3.5 Combination of masks

The values for the different thresholds in the combination of the 17 individual masks are depicted in Table 5.7. In both datasets, the best values are obtained with a threshold value of 7 or 8. The method can achieve simultaneously a sensitivity, specificity, precision and accuracy above 0.8 in both datasets.

Table 5.7: Sensitivity, specificity, accuracy, and precision for each threshold of the complete set of segmentation masks.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
thr	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
2	0.951	0.422	0.716	0.683	0.975	0.397	0.679	0.615	0.998	0.187	0.692	0.668
3	0.930	0.530	0.749	0.720	0.959	0.499	0.720	0.653	0.993	0.379	0.757	0.719
4	0.883	0.697	0.793	0.789	0.931	0.668	0.790	0.731	0.987	0.504	0.799	0.760
5	0.859	0.775	0.812	0.827	0.917	0.746	0.821	0.775	0.973	0.624	0.835	0.802
6	0.843	0.823	0.824	0.854	0.906	0.794	0.841	0.805	0.945	0.720	0.852	0.837
7	0.829	0.855	0.830	0.873	0.895	0.827	0.851	0.827	0.903	0.802	0.854	0.869
8	0.815	0.875	0.830	0.886	0.883	0.848	0.856	0.842	0.856	0.857	0.843	0.893
9	0.799	0.895	0.829	0.899	0.869	0.869	0.859	0.858	0.817	0.893	0.831	0.912
10	0.780	0.915	0.827	0.914	0.852	0.892	0.861	0.876	0.781	0.925	0.822	0.934
11	0.757	0.937	0.824	0.932	0.833	0.917	0.865	0.899	0.733	0.961	0.807	0.963
12	0.729	0.955	0.816	0.948	0.806	0.937	0.862	0.919	0.617	0.980	0.743	0.970
13	0.690	0.971	0.801	0.964	0.768	0.957	0.852	0.940	0.254	0.993	0.531	0.854
14	0.629	0.984	0.772	0.978	0.704	0.973	0.828	0.957	0.189	0.998	0.490	0.756
15	0.520	0.994	0.714	0.990	0.585	0.987	0.775	0.975	0.008	0.999	0.381	0.448
16	0.374	0.999	0.635	0.977	0.422	0.995	0.700	0.966	0.005	0.999	0.379	0.202

By looking at the graphics depicted in Fig. 5.27, it can be observed how, after the 8th threshold, any additional information has a detrimental effect on the results, as most statistical measures worsen. The values for VID_2 dataset are represented averaging both manual segmentations, while the third graphic represents the general value obtaining when averaging both datasets. This is also reinforced in the ROC curve (Fig. 5.28).

In view of the results, it is hinted that a smaller set of masks that obtains good results while improving the computation times may exist. To that end, the results from the individual masks were analysed in both subsets (Tables 5.3, 5.4, 5.5 and 5.6). The

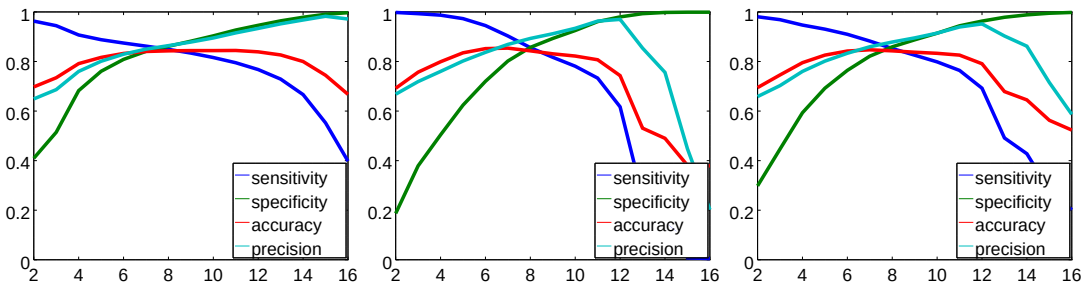


Figure 5.27: Evolution of sensitivity, specificity, accuracy and precision with the value of the threshold for the combination of the 17 segmentation masks in both datasets. From left to right: VID_2 dataset, IMG'_1 dataset, combination of both datasets.

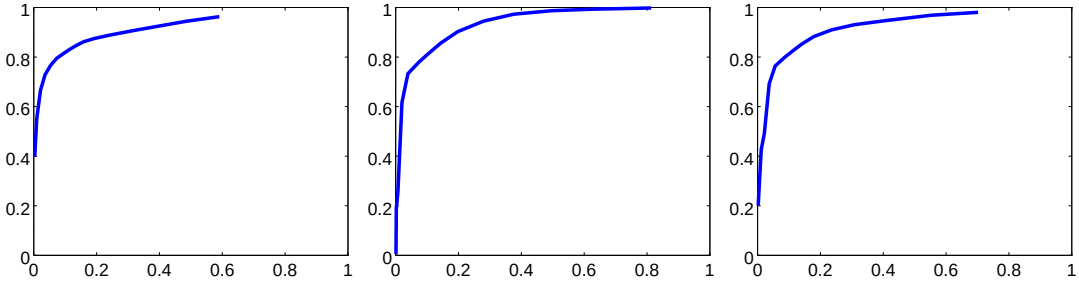


Figure 5.28: ROC curve for the combination of the 17 segmentation masks in both datasets. x-axis depicts the false positive rate and y-axis, the true positive rate. From left to right: VID_2 dataset, IMG'_1 dataset, combination of both datasets.

shape-based approaches were removed, as they have an inconsistent behaviour when the reference points are not present. M_C was also removed due to its poor behaviour in the IMG'_1 dataset. Thus, the reduced set consist of the 10 following masks: M_{TG}, $M_{TG'}$, M_{TS}, $M_{TS'}$, M_{TV}, M_{TL}, M_{MG}, M_{MO}, M_W and M_{SM}.

The results for the different thresholds in this reduced set are depicted in Table 5.8. In this test, the optimal values of threshold are again 7 and 8, as they achieve values above 0.8 in all parameters in both datasets. Between these two, the chosen value is 8, as the specificity is higher and, as it was detailed before, the most desirable characteristic is a low number of false positives.

The evolution of the parameters can be observed in Fig. 5.29, that graphically reinforces that the best threshold value is 8. Figure 5.30 depicts the ROC curve for

Table 5.8: Sensitivity, specificity, accuracy, and precision for each threshold of the combination of the reduced set of segmentation masks.

	VID_2 data set								IMG'_1 data set			
	Segmentation 1				Segmentation 2							
thr	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
2	0.951	0.422	0.716	0.683	0.975	0.397	0.679	0.615	0.998	0.187	0.692	0.668
3	0.930	0.530	0.749	0.720	0.959	0.499	0.720	0.653	0.993	0.379	0.757	0.719
4	0.883	0.697	0.793	0.789	0.931	0.668	0.790	0.731	0.987	0.504	0.799	0.760
5	0.859	0.775	0.812	0.827	0.917	0.746	0.821	0.775	0.973	0.624	0.835	0.802
6	0.843	0.823	0.824	0.854	0.906	0.794	0.841	0.805	0.945	0.720	0.852	0.837
7	0.829	0.856	0.830	0.873	0.895	0.827	0.851	0.827	0.903	0.802	0.854	0.869
8	0.815	0.875	0.830	0.886	0.883	0.848	0.856	0.842	0.856	0.857	0.843	0.893
9	0.799	0.895	0.829	0.899	0.869	0.869	0.859	0.858	0.817	0.893	0.831	0.912
10	0.780	0.915	0.827	0.914	0.852	0.892	0.861	0.876	0.781	0.925	0.822	0.934

this test.

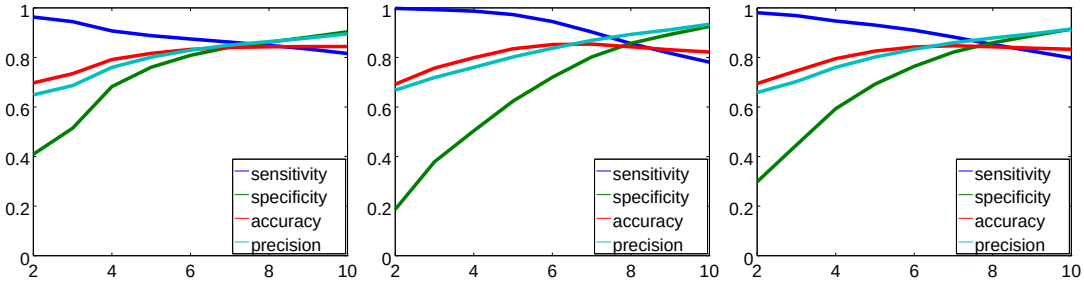


Figure 5.29: Evolution of sensitivity, specificity, accuracy and precision with the value of the threshold in the combination of the reduced set of segmentation masks in both datasets. From left to right: VID_2 dataset, IMG'_1 dataset, combination of both datasets.

It can be concluded that the approach that combines the reduced set of masks with a threshold $t = 8$ is the preferred choice, as it achieves good values in the four statistics while computing less masks, hence being significantly faster than the combination of the 17 approaches.

5.3.6 Enhancement techniques

The objective of these tests was to discover if pre- and post-processing techniques can improve the output of the segmentation approaches. To that end, the methods that

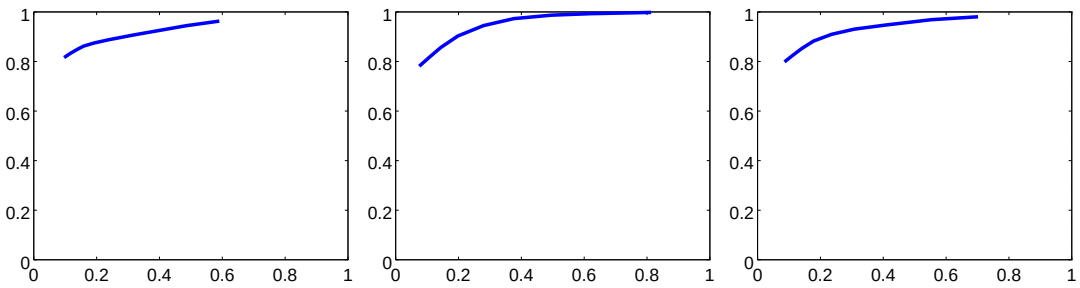


Figure 5.30: ROC curve for the combination of the reduced set of segmentation masks in both datasets. x-axis depicts the false positive rate and y-axis, the true positive rate. From left to right: VID_2 dataset, IMG'_1 dataset, combination of both datasets.

offered the best results in VID_2 dataset were chosen, and each technique was applied. The results obtained with the colour constancy techniques are depicted in Table 5.9.

Table 5.9: Sensitivity, specificity, accuracy, and precision for each colour constancy method applied before each segmentation procedure in the VID_2 dataset.

		Segmentation 1				Segmentation 2			
Mask	Method	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_{S_1}	CC_{GW}	0.761	0.860	0.793	0.869	0.831	0.846	0.829	0.834
	CC_{WP}	0.749	0.882	0.801	0.877	0.819	0.864	0.838	0.838
	CC_{WP_t}	0.775	0.803	0.780	0.811	0.830	0.781	0.802	0.757
	none	0.722	0.887	0.789	0.871	0.790	0.870	0.828	0.832
M_{ET}	CC_{GW}	0.741	0.913	0.804	0.909	0.798	0.884	0.831	0.862
	CC_{WP}	0.756	0.937	0.822	0.935	0.821	0.909	0.853	0.892
	CC_{WP_t}	0.854	0.871	0.848	0.888	0.907	0.827	0.854	0.830
	none	0.759	0.934	0.824	0.933	0.823	0.904	0.853	0.889
M_C	CC_{GW}	0.849	0.803	0.821	0.857	0.915	0.777	0.842	0.814
	CC_{WP}	0.840	0.841	0.830	0.871	0.910	0.814	0.856	0.830
	CC_{WP_t}	0.920	0.699	0.817	0.800	0.977	0.667	0.819	0.747
	none	0.837	0.842	0.829	0.869	0.906	0.815	0.854	0.827
M_{TV}	CC_{GW}	0.758	0.882	0.800	0.883	0.828	0.863	0.833	0.846
	CC_{WP}	0.799	0.833	0.801	0.852	0.868	0.817	0.830	0.814
	CC_{WP_t}	0.870	0.742	0.802	0.806	0.932	0.721	0.815	0.760
	none	0.801	0.830	0.801	0.850	0.870	0.813	0.829	0.812
M_{SM}	CC_{GW}	0.725	0.938	0.810	0.932	0.793	0.916	0.849	0.896
	CC_{WP}	0.737	0.926	0.811	0.922	0.804	0.905	0.848	0.886
	CC_{WP_t}	0.842	0.825	0.824	0.857	0.903	0.795	0.841	0.808
	none	0.818	0.829	0.814	0.854	0.877	0.800	0.832	0.805

CC_{WP_t} has the effect of lowering specificity, an undesirable consequence. The same applies to CC_{GW}, except for M_{TV} and M_{SM}, that are slightly improved. Finally, the effect of CC_{WP} is almost unnoticeable, with the exception of M_{SM}, that noticeably increases specificity, but at the cost of the same value in sensitivity. As the sensitivity values are too low after this, it is not worth it.

Regarding the filtering algorithms, the results are depicted in Table 5.10. Some of the approaches experiment a slight improvement. However, the results are not too noticeable in any of the methods.

Table 5.10: Sensitivity, specificity, accuracy, and precision for each filter applied before each segmentation procedure in the VID_2 dataset.

		Segmentation 1				Segmentation 2			
Mask	Method	Sens.	Spec.	Acc.	Prec.	Sens.	Spec.	Acc.	Prec.
M_{S_1}	F_G	0.777	0.835	0.797	0.857	0.830	0.804	0.815	0.799
	F_M	0.784	0.818	0.794	0.833	0.832	0.786	0.807	0.770
	F_B	0.783	0.833	0.801	0.854	0.837	0.803	0.818	0.799
	F_W	0.728	0.837	0.763	0.824	0.775	0.810	0.785	0.769
	none	0.722	0.887	0.789	0.871	0.790	0.870	0.828	0.832
M_{ET}	F_G	0.783	0.928	0.835	0.929	0.848	0.896	0.860	0.883
	F_M	0.790	0.922	0.836	0.924	0.852	0.888	0.858	0.876
	F_B	0.781	0.930	0.834	0.930	0.844	0.897	0.859	0.884
	F_W	0.619	0.939	0.746	0.927	0.670	0.916	0.782	0.883
	none	0.759	0.934	0.824	0.933	0.823	0.904	0.853	0.889
M_C	F_G	0.883	0.760	0.823	0.837	0.935	0.730	0.824	0.786
	F_M	0.880	0.731	0.807	0.821	0.937	0.702	0.814	0.770
	F_B	0.870	0.771	0.818	0.839	0.930	0.743	0.830	0.792
	F_W	0.864	0.817	0.835	0.857	0.929	0.789	0.851	0.812
	none	0.837	0.842	0.829	0.869	0.906	0.815	0.854	0.827
M_{TV}	F_G	0.795	0.839	0.801	0.856	0.866	0.823	0.833	0.819
	F_M	0.795	0.843	0.803	0.859	0.866	0.827	0.834	0.822
	F_B	0.795	0.839	0.802	0.856	0.867	0.823	0.833	0.819
	F_W	0.793	0.838	0.800	0.854	0.863	0.821	0.831	0.817
	none	0.801	0.830	0.801	0.850	0.870	0.813	0.829	0.812
M_{SM}	F_G	0.736	0.924	0.809	0.919	0.805	0.904	0.848	0.885
	F_M	0.735	0.927	0.810	0.922	0.804	0.906	0.849	0.887
	F_B	0.735	0.925	0.809	0.920	0.804	0.905	0.848	0.886
	F_W	0.733	0.928	0.809	0.923	0.801	0.907	0.848	0.888
	none	0.818	0.829	0.814	0.854	0.877	0.800	0.832	0.805

Therefore, both tests (filtering and colour constancy) were not validated with the second image set, because the results obtained during development with VID_2 dataset were subpar, as the computation times were increased with virtually no benefits.

The results for the shine removal procedure on the VID_2 dataset are depicted in Table 5.11. For this test, two manual segmentations were made, due to the difficulty of the task. As a human expert is unable to identify all the shining spots in the image, some of the validation parameters seem to indicate that the method performs poorly, even if it is not the case. Moreover, as this algorithm works with tiny areas in the image, the differences are magnified.

Table 5.11: Sensitivity, specificity, accuracy, and precision for the shine removal procedure.

Segmentation	Sensitivity	Specificity	Accuracy	Precision
1	0.577	0.997	0.997	0.217
2	0.552	0.998	0.997	0.216

The sensitivity values show that the number of false negatives is high, as there are regions defined in the manual segmentation that include areas that are not bright. However, the specificity is near perfect, as the region selected in the automatic approach is smaller, and the false positive rate is low.

5.4 Conclusions

In this chapter, several approaches to the segmentation of the bulbar conjunctiva are studied. This is one of the first steps in the computation of bulbar hyperaemia, and bears a special relevance, as a correct determination of the region of interest will guarantee an adequate environment for the next steps of the methodology. A total of 17 approaches were tested, both state-of-art and ad hoc, based on different image processing techniques. Moreover, combination of several individual masks, as well as pre- and post-processing techniques were analysed in order to further improve the results.

Given the differences in the images, there is not an optimal approach. None of the individual segmentation algorithms obtains optimal values for both datasets at once, even with pre-processing. However, the combination of 10 simple masks achieves a

value above 0.8 for specificity, sensitivity, accuracy and precision in both datasets. In view of the region of interest that the optometrists define, and taking into account the error that is derived of the existence of several possible optimal segmentations, these values will ensure that the segmentation is good enough for this environment.

Chapter 6

Extracting information from the images

Modelling the expert knowledge is the first obstacle to automate the grading process. As it was mentioned in Chapter 1, while qualitative descriptions on hyperaemia exist, it is not straightforward to apply them to an automatic system. One of the main issues is that there are several parameters that specialists take into account when evaluating hyperaemia, and the automatic system must cover all of them. Moreover, describing these parameters objectively and pointing out which ones are the most important are difficult tasks even for trained professionals. For example, the amount of red value in the image implies a higher hyperaemia level, but it does not have the same relevance if it is closer to the centre or the border of the eye, nor if the redness is caused by the vessels or the background of the conjunctiva.

Some examples of image characteristics related with bulbar hyperaemia are the general hue of the conjunctiva (Fig. 6.1, top) or the quantity of vessels (Fig. 6.1, bottom). A red or yellow tonality can suggest the presence of hyperaemia, while an almost white conjunctiva implies lower levels of the parameter. Regarding the disposition of the vessels, the more vessels, the higher the hyperaemia level, as the vessel engorgement can highlight even the thinner ones.

In order to include all the variables that optometrists analyse to grade hyperaemia, several image features have been computed in this work. Two divisions can be made of

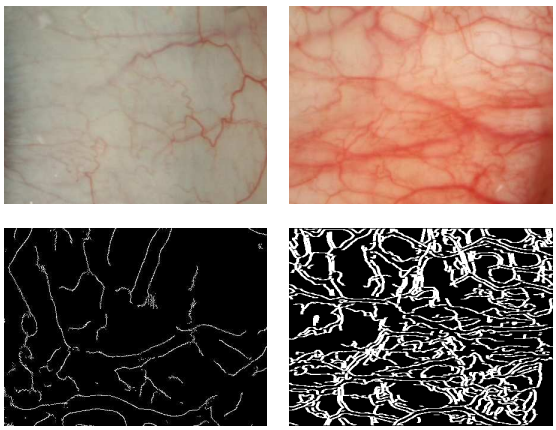


Figure 6.1: Image characteristics related with bulbar hyperaemia. Top: differences of hue in the bulbar conjunctiva. Bottom: differences in the quantity of vessels in the bulbar conjunctiva.

these features. First, depending on which parts of the region of interest are included in the computation, they can be divided as features that are computed only in the vessels, features that are computed only in the background and features that are computed in the whole conjunctiva. Second, the features can be divided taking into account if the information is only relative to vessel shape or disposition, or if they include hue information.

This chapter reviews the third step of the automatic methodology for bulbar hyperaemia grading, that involves the computation of image features, the comparison between each feature and the experts' evaluations, the analysis of the interaction between features and, finally, several approaches to combine the features.

6.1 Definition of the image features

In order to restrict the computation to certain elements of the image (background, vessels or whole conjunctiva), a Canny edge extraction algorithm [66] is used to highlight the conjunctival vessels. This algorithm applies a gaussian operator to smooth the image and then highlights the regions with a high first derivative, which will be the edges. The output of the algorithm is used to apply a mask that depends on which part of the image is being analysed, as depicted in Fig. 6.2. The performance of the Canny edge detection algorithm was evaluated in 106 manually segmented vessels. 94% of the

vessels were correctly extracted by the algorithm. In this experiment, the standard deviation of the gaussian operator was set to 9 while the low and high thresholds that establish the margins to track the ridges were set to 0.5 and 0.55 respectively. From this point forward, the edge image is labelled E; the pixels belonging to the edges, VE; and the non-edge pixels inside the conjunctiva region, $\overline{VE}$.

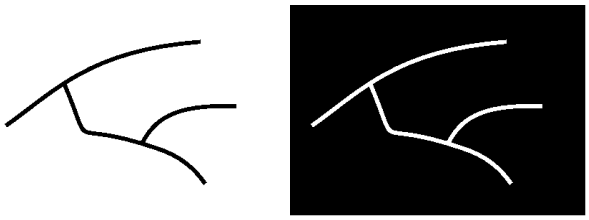


Figure 6.2: Different areas used for the computation. Left: mask for the features that use only the background. Right: mask for the features that use only the vessels.

A total of 25 image features were computed. They were chosen from the literature [37, 2] as well as following the suggestions of optometrists. As there is a limited number of relevant attributes in the conjunctiva, some of the features are highly related or computed in a similar manner. However, it is necessary to reflect all the possibilities in order to prevent any information loss. The redundancy that is expected to appear can be minimised in a later stage of the process.

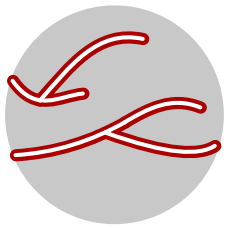
In all the equations that are depicted in this section, the input of the feature computation stage is an image I, where the conjunctiva has been already segmented. Therefore, the letters n and m represent the number of rows and columns of the image, but restricted to the conjunctiva, as the points taken into account are within this area. Thus, $n \times m$ represents the size of the conjunctiva and not the whole image. The variables i and j refer to a given pixel's row and column within the conjunctiva.

Table 6.1 shows the features that only use information regarding vessel quantity or width, labelled with a capital letter plus the subindex v. Three of the proposed approaches measure vessel quantity. While features A_v and P_v take into account the vessels in the whole conjunctiva, the feature C_v counts the number of vessels, but restricting the computation to n_r rows of the image separated a *step*. In the equation, M is a mask defined as:

$$M_{ij} = \begin{cases} 0 & i \bmod step \neq 0 \\ 1 & i \bmod step = 0 \end{cases}$$

The last vessel-related feature, W_v, measures the average width of the vessels. To that end, a series of circumferences centred in the centre of the image with varying radii r are created. The points where the circumferences cut a vessel are located. Next, the width of those vessels at the cutoff points is computed by means of an active contour algorithm [67]. In the equations, W defines the width values for the considered points, and r ranges from $n/2 * c$ to $n/2$ where c is the number of circumferences.

Table 6.1: Image features that compute vessel quantity or width.

 X_v	Vessel count (C_v)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} E_{ij}M_{ij}}{n_r}$
	Vessel occupied area (A_v)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} VE_{ij}}{nm}$
	Percentage of vessels (P_v)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} VE_{ij}}{nm}100$
	Vessel width (W_v)	$\frac{\sum_{r=1}^{\rho}\sum_{c=1}^{\kappa} W_{rc}}{\rho\kappa}$

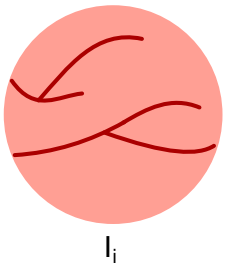
Some features require parameter tuning. In these cases, the values were empirically chosen. The parameters for the feature C_v are $n_r = 10$ and $step = 10$. The value selected for feature W_v is $c = 10$.

There are three groups of features that take hue into account. To compute them, there are several parameters calculated in different colourspaces in order to find which ones provide a better insight on the experts' perception. Usually, each channel of a given colourspace that is included in the computation is processed separately. R, G and B represent each channel's values in RGB colourspace; H, S and V, in HSV colourspace; and L, a and b, in L*a*b* colourspace. In order to obtain the values for the red hue in the HSV colourspace, a correction of $|H - 128|$ is applied to the H channel. This is caused by the definition of that channel, as the red colour is centred at 0 and, therefore, the closer to zero (or 255, as the definition is circular), the redder the pixel. Regarding the a channel from L*a*b*, positive values imply a red hue, and negative ones, green.

Table 6.2 shows the features that take into account the hue in the whole conjunctiva,

labelled with a capital I followed by a numeric subscript.

Table 6.2: Image features that compute the hue in the whole conjunctiva.

 I_i	Relative image redness (I_1)	$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{R_{ij}}{R_{ij}+G_{ij}+B_{ij}}\right)$
	Difference red-green of the image (I_2)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}-G_{ij})}{nm}$
	Difference red-blue of the image (I_3)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}-B_{ij})}{nm}$
	Red hue value (I_4)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\lvert 128-H_{ij}\rvert}{nm}$
	L*a*b* a-channel of the image (I_5)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}}{nm}$

Next, Table 6.3 lists the features that take into account the hue in the vessels, labelled with a capital V followed by a numeric subscript. The feature V_6 computes the red hue value taking into account not only the current pixel but also its neighbouring ones. μ is the value for the neighbourhood, computed as:

$$\mu_{ij} = \frac{\sum_{k=-s/2}^{s/2}\sum_{l=-s/2}^{s/2}\overline{VE}_{ij}Hi+k,j+l}{s^2}$$

where s is the size of the considered window. In this work, the chosen value was $s = 3$.

Finally, Table 6.4 shows the features that take into account the hue in the background of the conjunctiva, labelled with a capital B plus a numeric subscript.

6.2 Experts' evaluations vs image features

In order to mimic the specialists' behaviour, the system must be able to produce hyperaemia evaluations in a given grading scale. To that end, the first step is to analyse both the manual and automatic values in order to find a relationship between them.

For the sake of clearness, the pairwise correlation between features was computed previously, in order to *group* the features that are related. Fig. 6.3 shows the values for VID dataset. This data set was selected because it has evaluations in two grading scales, and it is reasonable to assume that the most relevant features may vary, in view of the differences in the images that each scale uses as prototypes. The correlation

Table 6.3: Image features that compute the hue in the vessels.

Relative vessel redness (V_1)	$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{R_{ij}VE_{ij}}{R_{ij}+G_{ij}+B_{ij}}\right)$
Difference red-green in vessels (V_2)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}-G_{ij})VE_{ij})}{nm}$
Difference red-blue in vessels (V_3)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}-B_{ij})VE_{ij})}{nm}$
Percentage of red (RGB) (V_4)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}VE_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}100$
Percentage of red (HSV) (V_5)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}H_{ij}VE_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}VE_{ij}}100$
Redness with neighbourhood (V_6)	$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{H_{ij}VE_{ij}}{\mu_{ij}}$
L*a*b* a-channel in vessels (V_7)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij}VE_{ij})}{nm}$

Table 6.4: Image features that compute the hue in the background.

Yellow in background (RGB) (B_1)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}+G_{ij})\overline{VE}_{ij})}{nm}$
Yellow in background (HSV) (B_2)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\lvert 240-H_{ij}\rvert\overline{VE}_{ij})}{nm}$
Yellow in background (L*a*b*) (B_3)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(b_{ij}\overline{VE}_{ij})}{nm}$
Red in background (RGB) (B_4)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}\overline{VE}_{ij})}{nm}$
Red in background (HSV) (B_5)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\lvert 128-H_{ij}\rvert\overline{VE}_{ij})}{nm}$
Red in background (L*a*b*) (B_6)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij}\overline{VE}_{ij})}{nm}$
White in background (RGB) (B_7)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((R_{ij}+G_{ij}+B_{ij})\overline{VE}_{ij})}{nm}$
White in background (HSV) (B_8)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}((V_{ij}+S_{ij})\overline{VE}_{ij})}{nm}$
White in background (L*a*b*) (B_9)	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(L_{ij}\overline{VE}_{ij})}{nm}$

was computed using an iteratively reweighted least squares algorithm with a bisquare weighting function [68].

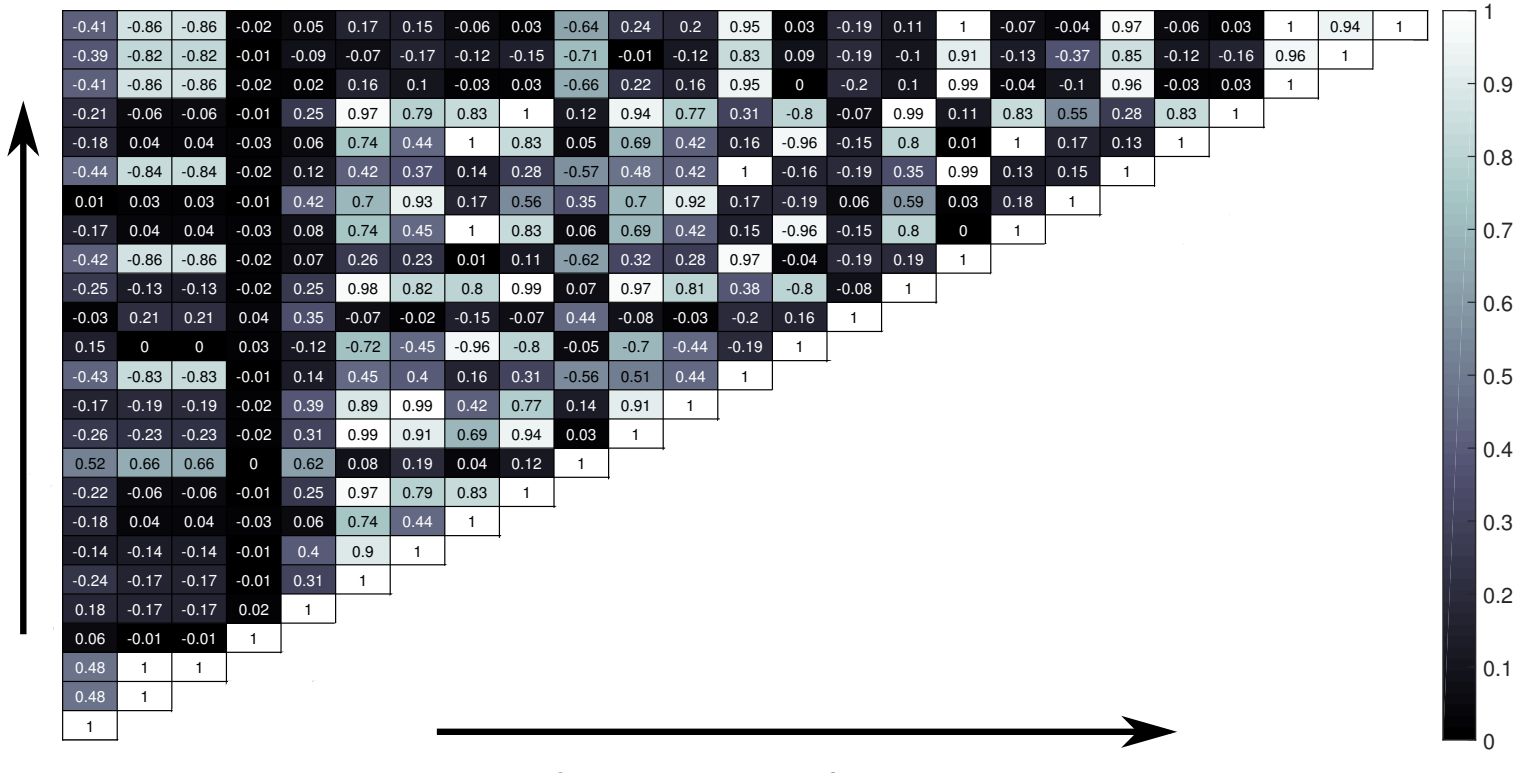


Figure 6.3: Pairwise feature correlation for VID dataset. Both axis represent the 25 features in the order that they were defined, placed from bottom to top and from left to right.

Table 6.5 shows the groups of features that have an absolute correlation of at least 0.7. This value was chosen in order to depict the groups of features that have a strong correlation among them. The kappa index and correlation between feature and expert was computed only for one characteristic of each group, as if two features have a high correlation between them, they will have a similar correlation with the experts' evaluation.

Table 6.5: Groups of features in the VID dataset.

C_v
A_v P_v V_4 B_1 B_4 B_7 B_8
W_v
I_1
I_2 I_3 I_4 I_5 V_2 V_3 V_5 V_7 B_2 B_3 B_5 B_6
V_1 B_9
V_6

The kappa index was computed for the feature-expert comparison. To that end, the range of values of each feature was trasnformed to the grading scales ones, this is, 0 to 4 and 1 to 4 for the Efron and BHVI scales, respectively. As in Chapter 2, *po* and *pe* represent the observed and random agreement, respectively. The null hypothesis H_0 represents that the observed agreement is accidental. The significance level α is 0.05. The agreement is displayed in a scale from 0 to 5, from lower to higher. Tables 6.6 and 6.7 depict the values for the VID dataset in the Efron and the BHVI scales respectively. The average of the experts' measurements was used. The column *step* represents the class division that was performed: one decimal value, integer and half integer and only integer values.

The agreement was generally poor or slight, even with the wider step. This hints at that there is not a direct relationship between each feature and the experts' evaluation. In some of the cases, Cohen's kappa falls below 0, indicating no agreement at all. The best agreement is obtained by I_1, V_2 and I_4. The Efron scale shows lower agreement on average than the BHVI scale, as only the test with V_2 and a step of 0.5 rejects the null hypothesis. Therefore, it is expected that to find a relationship between the image features and the experts values will be easier in the BHVI than in the Efron scale.

Table 6.6: Cohen's kappa coefficient for the evaluation of experts E_1 and E_2 (two evaluations each expert) compared with image features (Efron scale).

step	feature	po	pe	kappa	agreement	var	p	H_0
0.1	C_v	0.0095	0.0441	-0.0362	0	0.0005	0.1006	Accept
	A_v	0.0190	0.0427	-0.0247	0	0.0005	0.2484	Accept
	V_1	0.0667	0.0485	0.0191	1	0.0005	0.3995	Accept
	I_1	0.0667	0.0493	0.0182	1	0.0005	0.4280	Accept
	V_2	0.0571	0.0465	0.0111	1	0.0005	0.6205	Accept
	I_4	0.0571	0.0404	0.0175	1	0.0005	0.4215	Accept
	V_6	0.0000	0.0092	-0.0092	0	0.0009	0.7636	Accept
	W_v	0.0381	0.0243	0.0141	1	0.0004	0.4950	Accept
0.5	C_v	0.1429	0.2139	-0.0903	0	0.0032	0.1109	Accept
	A_v	0.1524	0.2148	-0.0795	0	0.0029	0.1399	Accept
	V_1	0.2952	0.2351	0.0786	1	0.0034	0.1755	Accept
	I_1	0.2286	0.2540	-0.0340	0	0.0033	0.5511	Accept
	V_2	0.3238	0.2299	0.1219	1	0.0035	0.0401	Reject
	I_4	0.2381	0.1986	0.0492	1	0.0038	0.4221	Accept
	V_6	0.0286	0.0473	-0.0196	0	0.0066	0.8090	Accept
	W_v	0.1238	0.1173	0.0074	1	0.0052	0.9181	Accept
1.0	C_v	0.3905	0.3846	0.0096	1	0.0082	0.9157	Accept
	A_v	0.3714	0.3814	-0.0161	0	0.0072	0.8496	Accept
	V_1	0.4667	0.3886	0.1277	1	0.0068	0.1212	Accept
	I_1	0.4571	0.4519	0.0096	1	0.0063	0.9041	Accept
	V_2	0.4381	0.4101	0.0475	1	0.0084	0.6035	Accept
	I_4	0.4857	0.4073	0.1324	1	0.0081	0.1425	Accept
	V_6	0.0667	0.0956	-0.0320	0	0.0147	0.7922	Accept
	W_v	0.2286	0.2277	0.0012	1	0.0185	0.9931	Accept

Table 6.7: Cohen's kappa coefficient for the evaluation of experts E_1 and E_2 (two evaluations each expert) compared with image features (BHVI scale).

step	feature	po	pe	kappa	agreement	var	p	H_0
0.1	C_v	0.0381	0.0571	-0.0202	0	0.0006	0.4242	Accept
	A_v	0.0476	0.0540	-0.0067	0	0.0006	0.7816	Accept
	V_1	0.0762	0.0610	0.0162	1	0.0007	0.5322	Accept
	I_1	0.0952	0.0659	0.0315	1	0.0007	0.2334	Accept
	V_2	0.0476	0.0561	-0.0090	0	0.0006	0.7176	Accept
	I_4	0.0667	0.0621	0.0048	1	0.0007	0.8544	Accept
	V_6	0.0000	0.0120	-0.0121	0	0.0012	0.7270	Accept
	W_v	0.0286	0.0455	-0.0178	0	0.0007	0.4959	Accept
0.5	C_v	0.3048	0.3030	0.0025	1	0.0042	0.9696	Accept
	A_v	0.2190	0.2584	-0.0531	0	0.0041	0.4076	Accept
	V_1	0.3619	0.2905	0.1006	1	0.0044	0.1302	Accept
	I_1	0.3810	0.3294	0.0768	1	0.0041	0.2278	Accept
	V_2	0.3429	0.2644	0.1067	1	0.0052	0.1390	Accept
	I_4	0.4095	0.3009	0.1554	1	0.0054	0.0352	Reject
	V_6	0.0381	0.0550	-0.0179	0	0.0092	0.8522	Accept
	W_v	0.2000	0.2203	-0.0261	0	0.0093	0.7869	Accept
1.0	C_v	0.4286	0.4680	-0.0742	0	0.0116	0.4902	Accept
	A_v	0.5048	0.4642	0.0757	1	0.0089	0.4225	Accept
	V_1	0.5048	0.4767	0.0536	1	0.0083	0.5577	Accept
	I_1	0.6000	0.4792	0.2320	2	0.0092	0.0153	Reject
	V_2	0.6476	0.4650	0.3413	2	0.0135	0.0033	Reject
	I_4	0.5810	0.4740	0.2033	1	0.0136	0.0817	Accept
	V_6	0.0952	0.1088	-0.0153	0	0.0156	0.9026	Accept
	W_v	0.4857	0.4746	0.0212	1	0.0254	0.8939	Accept

6.3 Feature selection

As it is not possible to establish a clear relationship between the experts' evaluations and the image features, the next step is to combine the features, expecting that the combined information will bring both the manual and the automatic approaches closer. However, as some of the proposed features use similar image information, it is reasonable to think that there may be redundant information if they are stacked together. Therefore, it is important to analyse the relations among features, in order to discard the ones that do not provide useful information, as they can be a source of noise and/or bias.

By observing the results of the pairwise correlation for both groups (Fig. 6.3), it is clear that there is low correlation among most of the features. Moreover, some of the features that are expected to be related by observing the equations, present a correlation close to zero, such as (I_1, I_4, I_5) or (B_1, B_2, B_3). Also, some features that seem unrelated have a high correlation, such as I_2 with V_7. This supports the necessity of implementing as many features as possible, even if they may seem redundant.

As most of the features have no apparent relation to the values of the manual evaluations, the next step is to analyse if combining them the results could be improved. The first option is to combine all the 25 features. However, as there is redundancy in the feature set, a better choice may be to apply a procedure to select the features that provide the most information, while removing the redundant ones.

There are two main options to tackle this problem: feature selection and feature extraction. Feature extraction techniques build a new set of features from the original feature set. They are typically used in dimensionality reduction problems, and they involve a transformation of the features. This transformation can be non reversible, thus producing loss of information. Some examples of feature extraction are Principal Component Analysis (PCA) [69] or Linear Discriminant Analysis (LDA) [70].

Feature selection techniques [71] examine the original set of features in order to obtain a subset which contains a lower number of features while preserving most of the information. In order to decide if each feature is worth including or not, they use a certain criteria, such as correlation or information gain. These techniques do not involve a transformation of the original data and, therefore, are more useful in this

scenario, as they can help to gain a better understanding of which features are the most relevant for the specialists. The objective of feature selection techniques is to obtain the features that are *relevant*, namely those that vary systematically with category membership [72]. Feature selection techniques can be divided into three groups:

Filters are the fastest ones, as they evaluate general characteristics of the data, such as the correlation. These techniques do not need learning or the construction of a model in order to perform the selection.

Wrappers use a model or prediction system in order to form the feature subsets. As a direct consequence, they tend to be slower, but also to provide more accurate results [73]. Wrappers tackle the problem as a search problem, creating and evaluating several feature combinations and, finally, choosing the best one. The search strategy may vary.

Embedded methods blend the feature selection with the training process of the prediction model. Like the wrappers, these methods work altogether with the prediction model. Therefore, they usually offer precise results, but have the disadvantage of the highest computational cost of the three groups.

In this work, five approaches were tested:

- Two filters, correlation-based feature selection (CFS) and Relief.
- Two wrappers, one based on M5 model tree (M5) and another based on support vector machines with the sequential minimal optimisation method (SMOReg).
- An embedded method, also based on support vector regression (SVR-RFE).

In addition, as a cross-validation technique was used, the methods can output different subsets for each fold. Therefore, it is necessary to define a criteria to perform the final selection of the feature set. Further details regarding cross-validation techniques can be found in Appendix E.

The next sections explain the process in further detail with the VID_1 dataset. This dataset was chosen because it is labelled in both Efron and BHVI and, moreover, two

of the experts labelled it twice. Finally, an evaluation with the IMG_1 dataset is also included.

6.3.1 CFS

Correlation based Feature Selection [74] is a filter method and, therefore, independent from the learning method. Its original focus were the classification problems. Therefore, in order to apply it to a regression scenario, a previous discretisation stage is needed to transform the data. To that end, the algorithm MDL [75] was used. CFS searches for features that are highly correlated with the problem while maintaining a low correlation with each other. The output of the method is a subset containing the relevant features. The method was tested with and without normalisation, and the subsets depicted in Table 6.8 were obtained.

Table 6.8: Feature subset for each fold using CFS in VID_1 dataset.

k	Efron	Efron normalised	BHVI	BHVI normalised
1	I_5, B_6, B_8, W_v	I_1, I_5, B_5, B_6, W_v	V_6, I_5, B_6, W_v	I_1, I_5, B_5, B_6, W_v
2	I_1, I_5, B_6, B_8, W_v	I_1, V_5, I_5, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, V_5, I_5, B_6, W_v
3	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v
4	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, B_6, B_8, W_v	I_1, V_6, I_5, V_7, B_2, B_6
5	I_1, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_9, W_v
6	A_v, I_1, B_6, B_8, W_v	V_6, I_5, V_7, B_6, B_9, W_v	I_1, I_5, B_6, B_8, W_v	V_6, I_5, V_7, B_6, B_9, W_v
7	I_1, I_5, B_2, B_6, B_8	I_1, V_5, I_5, B_2, B_6, B_8	I_1, I_5, B_2, B_6, B_8	I_1, V_5, I_5, B_2, B_6
8	I_1, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v
9	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_9, W_v
10	A_v, I_1, I_5, B_6, B_8, W_v	I_1, I_5, V_7, B_2, B_6, B_9, W_v	A_v, I_1, I_5, B_6, B_8, W_v	I_1, I_5, B_2, B_6, B_8, W_v

6.3.2 Relief

Relief [76, 77] is another filter method, that belongs to the sub type known as *ranker methods*. Instead of constructing a subset of features, rankers sort them by using a goodness metric, from the most relevant to the least. The output of the method is the whole list of features sorted by relevance. Therefore, there is an additional step to decide which features are being considered. To this end, there are several approaches, from establishing a number of features to be considered, such as the best n of each fold,

to setting a relevance threshold and considering only the features that score a higher value. In this work, the latter was used, so only the features that scored higher than the value determined by the following equation are allowed:

$$thr = \frac{1}{2N} \sum_{k=1}^{N} \max(g(x_k)) \tag{6.1}$$

where N is the number of folds, x_k is the feature in the fold k and g is the gain function. The gain function maps each feature x in the current fold k to the relevance value calculated by the Relief method. Thus, g is directly proportional to the relevance of the feature and, therefore, to the gain. The selected features are depicted in Table 6.9. The results are the same in both normalised and raw values, with minor differences in the positions. Thus, the table does not makes this distinction.

Table 6.9: Feature order for each fold using Relief in VID_1 dataset.

k	Efron	BHVI
1	V_6, B_6, I_5, V_5, V_7	V_6, P_v, A_v, V_5, B_6, I_5, I_4, B_5, B_2
2	V_6, V_5, V_7, B_6, I_5	V_6, V_5, P_v, A_v, V_7, I_1, B_6, I_5, I_4, B_5, B_2
3	V_6, B_6, I_5, V_7, V_1, V_5, V_2, I_2	V_6, B_6, I_5, V_1, V_7, V_5, I_1, V_2, I_2
4	B_6, I_5, V_7, V_6, I_2, V_2, V_5, V_4, B_9, B_1, B_7	B_6, I_5, V_4, B_9, V_6, B_1, V_7, P_v, A_v, B_7, B_4, I_2, V_5, V_2
5	V_6, P_v, A_v	V_6, P_v, A_v, B_9, B_7, B_1
6	B_6, I_5, V_7	B_6, I_5, V_7, V_5
7	V_6, V_5	V_6, P_v, A_v, V_5
8	V_6, V_7, I_5, B_6, I_1, V_1	V_6, I_1, P_v, A_v, B_6, I_5, V_1, V_7
9	V_6, V_7, I_5, B_6, V_2	V_6, P_v, A_v, B_6, I_5, V_5, V_7
10	V_6, V_5, B_6, I_5, V_7	V_6, P_v, A_v, V_5, B_6, I_5, V_7

6.3.3 M5

The first wrapper applies the M5 algorithm [78, 79] for generating model trees that have linear regression functions in their nodes. The separate-and-conquer approach is used to build an M5 tree in each iteration. Then, the best leaf is made into a rule. The search strategy is best-first. The selected features are shown in Table 6.10.

Table 6.10: Feature subset for each fold using M5 in VID_1 dataset.

k	Efron	BHVI
1	V_1, V_4, V_7, B_6, B_7, W_v	C_v, I_3, V_4, V_6, B_3, B_4, B_6, B_7, B_8, B_9
2	I_1, I_5, V_7, B_2, B_6	V_6, V_7, B_6
3	V_6, V_7, B_6, B_8	I_1, V_6, V_7, B_6
4	I_1, V_2, I_3, B_3, B_6	C_v, I_2, V_6, B_6
5	C_v, I_4, B_6	I_4, V_6, B_6
6	V_2, V_3, I_4, V_6, B_6	I_1, V_3, V_6, B_6, B_7
7	B_6	V_1, I_2, V_5, V_6, B_4, B_6
8	A_v, V_6, I_5	A_v, I_2, V_6, B_6
9	I_1, I_2, I_5, B_9, W_v	I_1, I_3, V_4, V_7, B_4, B_6, W_v
10	C_v, I_4, V_6, B_6	C_v, I_2, V_6, I_5, V_7, B_4, B_6, B_7, B_8

6.3.4 SMOReg

The second wrapper is based on the support vector machine for regression (SVR). The algorithm includes the improvements proposed in [80] for the sequential minimal optimisation (SMO) method. The search strategy is also best-first. The selected features are depicted in Table 6.11.

Table 6.11: Feature subset for each fold using SMOReg in VID_1 dataset.

k	Efron	BHVI
1	C_v, V_1, I_1, V_2, I_2, V_4, I_5, B_1, B_3, B_4, B_6, B_7, B_9	V_6, B_3, B_6
2	I_1, B_6	I_1, V_3, I_3, I_5, B_6
3	C_v, I_1, V_2, I_5, B_6	V_2, V_6, I_5, B_6
4	C_v, V_1, I_1, V_4, I_5, B_1, B_3, B_4, B_6, B_7, B_9	C_v, I_2, V_3, V_6, B_3, B_6
5	I_1, V_4, B_1, B_4, B_6	I_1, B_3, B_6
6	V_2, V_6, I_5, B_6	V_3, I_3, V_6, I_5, B_6
7	C_v, I_1, V_5, I_5, B_6	C_v, A_v, V_4, V_5, V_6, I_5, B_1, B_4, B_6, B_9
8	C_v, I_1, V_3, I_5, B_6	C_v, V_3, V_6, I_5, B_6
9	V_1, I_1, I_3, V_4, I_5, B_1, B_4, B_6, B_7, B_9, W_v	C_v, I_1, B_3, B_6, W_v
10	C_v, V_1, I_5, B_6	C_v, V_6, I_5, B_3, B_6

6.3.5 SVR-RFE

The embedded method uses recursive feature elimination (RFE) with support vector regression [81]. It performs an iterative process that starts with the full set of features. The method assigns a weight to each feature. Features with the smallest absolute

weights are removed. The process continues until the minimum number of features, previously established, is reached. In this work, the minimum number of features was chosen empirically to be 10. Tests with smaller sets were performed, but the variability among folds was too high to draw conclusions on which were the preferred features overall. The selected features are depicted in Table 6.12.

Table 6.12: Feature subset for each fold using SVR-RFE in VID_1 dataset.

k	Efron	BHVI
1	A_v, I_4, P_v, V_5, V_6	A_v, I_4, V_5, V_6, W_v
2	A_v, I_3, V_4, V_5, B_4	A_v, I_3, V_4, V_5, B_4
3	A_v, V_5, V_6, B_5, W_v	A_v, V_5, V_6, B_5, W_v
4	A_v, I_3, I_4, V_5, V_6	A_v, I_3, I_4, V_5, I_5
5	A_v, I_3, V_4, V_5, B_4	A_v, I_3, V_4, V_5, B_4
6	A_v, V_1, I_3, I_4, V_5	A_v, I_3, V_5, V_6, W_v
7	A_v, I_3, V_4, V_5, I_5	A_v, I_3, V_4, V_5, I_5
8	A_v, I_3, I_4, V_5, B_4	A_v, I_3, V_5, V_6, W_v
9	A_v, V_1, I_4, V_5, V_6	A_v, V_1, I_4, V_5, V_6
10	A_v, I_4, P_v, V_5, I_5	C_v, V_1, I_4, P_v, V_6

6.4 Combination of features

Once the methods have outputted a result for each fold, the next step is to combine the results to obtain a general subset. There are several approaches to perform this task, depending on the number of features that were selected on each fold and the size of the differences among folds regarding the chosen features:

Intersection of features selected in each fold. The intersection set is not appropriate in this case, as there are methods where it would be empty.

Union of features selected in all the folds. This set is too inclusive, as the variability between folds is too big.

Threshold to establish a minimum number of folds were a feature must appear in order to be considered.

Figure 6.4 depicts the number of folds where each feature was chosen by each method. In view of the data, the chosen option was to establish a minimum number of folds for a feature to be considered. The threshold was empirically obtained, and fixed at 7. The selected features when considering only those that appear in at least 7 out of 10 folds are depicted in Table 6.13.

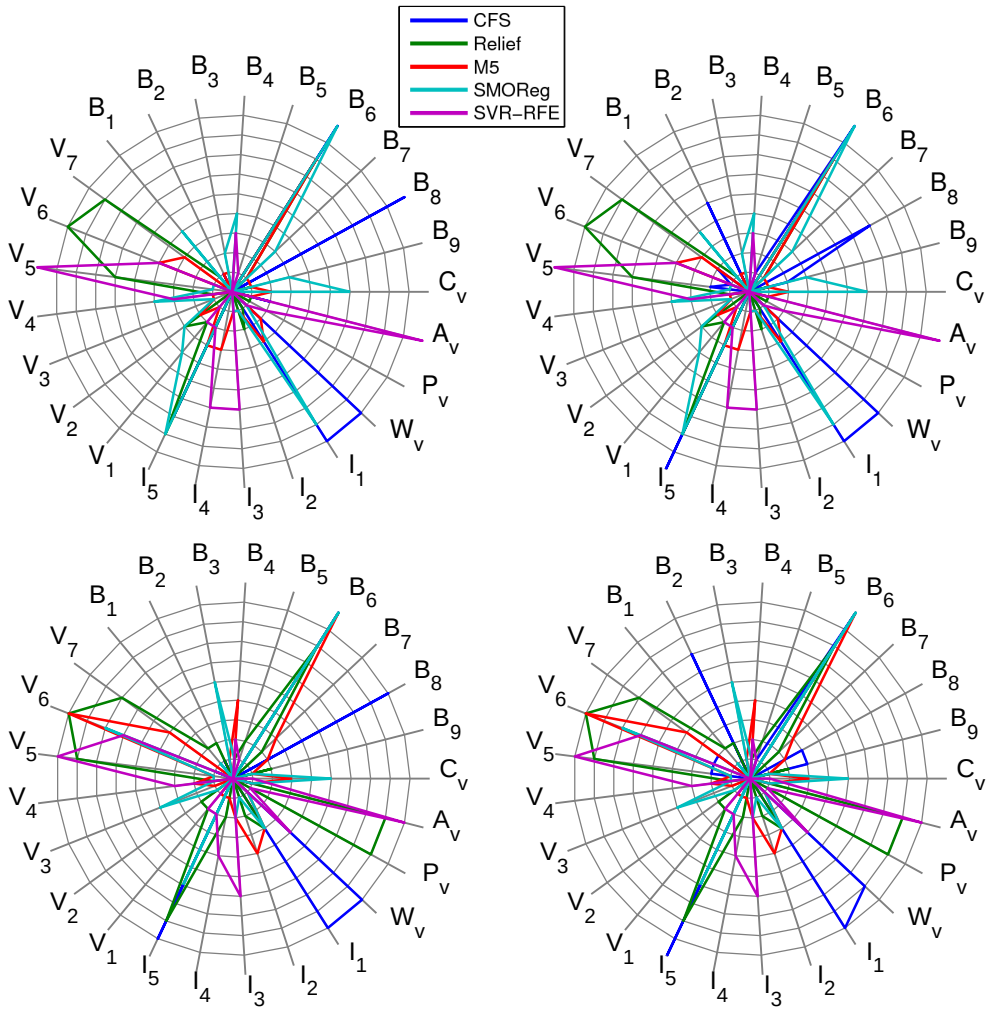


Figure 6.4: Plot depicting the number of folds where each feature was chosen. The centre of the plot represents that the feature is chosen in zero folds, and the outermost line, in all the ten folds. The top and bottom rows show the results in the Efron and the BHVI scale, respectively. Left: without normalisation. Right: normalised.

In view of the data, the differences between raw and normalised values are only noticeable in the ranker approach and, even then, they are minimal. Therefore, the

Table 6.13: Features that appear in at least 7 out of 10 folds in the VID_1 dataset.

Method	Scale	# selected features	Selected features
CFS	Efron	5	I_1, I_5, B_6, B_8, W_v
	BHVI	5	I_1, I_5, B_6, B_8, W_v
CFS (normalised)	Efron	5	I_1, I_5, B_6, B_8, W_v
	BHVI	5	I_1, I_5, B_2, B_6, W_v
Relief	Efron	4	V_6, I_5, V_7, B_6
	BHVI	7	A_v, P_v, V_5, V_6, I_5, V_7, B_6
M5	Efron	1	B_6
	BHVI	3	I_3, V_6, B_6
SMOReg	Efron	3	I_1, I_5, B_6
	BHVI	2	I_5, B_6
SVR-RFE	Efron	2	A_v, V_5
	BHVI	2	A_v, V_5

subsequent tests will not consider this difference. The most commonly chosen feature is B_6 in both grading scales, followed by I_5. CFS favours W_v, but it is the only method that chooses it. The background of the conjunctiva seems to be more discriminant than the vessels, and the colour-based features, preferred over the vessel-based ones.

Once several feature selection methods have been tested, a common approach is to combine the results of each method in order to create a general subset. The idea is to combine the strengths of the methods, since the features that have been selected by several approaches are likely to be the most relevant ones.

The situation is similar to the combination of the results obtained for the different folds, so the possible solutions are the aforementioned: the intersection or the union of the features selected by all the methods, or considering the features that appear in a minimum number of approaches. In this case, the selected option was to perform the union of the subsets, as the selection methods present a high variability. The final sets were the following:

- Efron: A_v, W_v, I_1, I_5, V_5, V_6, V_7, B_6, B_8.
- BHVI: A_v, P_v, W_v, I_1, I_3, I_5, V_5, V_6, V_7, B_2, B_6, B_8.

All the features that are selected in the Efron scale are also selected in the BHVI scale. For the latter, three additional features are selected: one that computes vessel

quantity (P_v), another that measures the yellow in the background (B_2) and a last one that computes the difference of red and blue in the image (I_3). If the features are divided in groups by following the criteria defined at the beginning of the chapter, on one hand the vessel-related and on the other hand each of the hue-related features in the three different areas (background, vessels and whole conjunctiva), all the groups are evenly represented. Also, there is not a clear consensus regarding the most relevant colour space. Regarding the hue, the red value is computed in all the features that compute the colour in the vessels (V_5, V_6 and V_7), and is also the most relevant in the whole image. The conjunctiva background presents more variability, as B_6 computes the red level, but B_8 measures the white.

It is interesting to note that the kappa coefficient results from section 6.2 concluded that the image features seemed to be closer to the experts' evaluations in the BHVI scale than in the Efron scale. However, all the feature selection techniques selected larger subsets in the BHVI experiments. Therefore, although the experts' evaluations in BHVI are closer to the selected individual features, the situation is different when taking into account several features at a time.

6.5 Local vs. global features

Several works on the matter, as well as the optometrists that evaluated the image sets, agree that not all the surface of the conjunctiva is given the same importance. That is, if an image feature such as a high level of redness in the vessels, takes place near the caruncle area, it is not as relevant as if it happens near the iris. Therefore, an analysis was performed dividing the eye in two regions as depicted in Fig. 6.5.

The same set of features was computed in the whole image and in both sides of the image: the iris side and the caruncle/corner of the eye side. The first step was to assess the differences in the automatic values for each area. Therefore, the pairwise correlation was computed for each feature in each pair of areas. The results depicted in Table 6.14 show how some features, such as I_2 or V_3, remain stable when they are computed in the whole image in comparison with one of the sides. However, other features show a lower correlation. Specifically, feature V_6 shows a poor correlation in

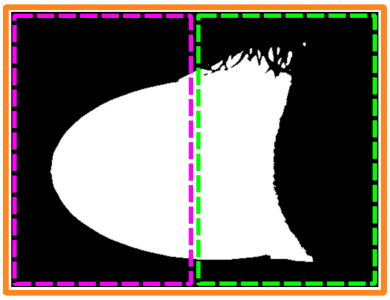


Figure 6.5: Sections of the image where the features are computed: whole image, iris side and corner of the eye side.

all the cases, and features V_1 and I_1 present low values when the corner of the eye is involved.

In order to observe how these differences affect the outputs of the automatic methodology, the feature selection approaches were applied to the features computed in each part of the image. The results are depicted in Table 6.15. Superscript G will refer to the features computed in the whole conjunctiva, I to those computed in the iris side, and C to those computed in the caruncle side.

In view of the results, there are some features that appear in most of the cases, such as V_7 and V_1. Regarding the area of computation, the table shows how most of the features are computed in the whole conjunctiva. However, several feature selection options take into account at least one feature computed for one of the sides. Moreover, half of the features selected by SVR-RFE and CFS are local features. This reinforces the idea that some characteristics have a higher relevance if they are computed in a given area. The percentage of selected local features is 50% for CFS (both scales), 38% and 17% for Relief, 33% and 0% for both wrappers, and 50% in SVR-RFE (both scales).

Finally, if the local feature selection results are compared with the global ones in Table 6.13, it becomes apparent that the vessels gain importance in the local approach. Most of the selected features are not the same. CFS selects I_5 and I_1 (BHVI only) but in the local regions. Relief chose V_6 and the two vessel-based A_v and P_v in BHVI. None of the wrappers have a single feature in common with the global-only experiment. Finally, SVR-RFE repeats only V_5 and, again, it does so in one of the sides of the eye.

Table 6.14: Average correlation between features computed in different areas of the eye in the VID dataset.

Feature	Global vs. iris side	Global vs. corner side	Iris side vs. corner side
C_v	0.708	0.726	0.027
A_v	0.812	0.779	0.282
P_v	0.812	0.779	0.282
W_v	0.033	-0.006	0.143
I_1	0.876	0.627	0.173
I_2	0.902	0.906	0.669
I_3	0.926	0.904	0.692
I_4	0.957	0.770	0.588
I_5	0.887	0.889	0.627
V_1	0.845	0.634	0.122
V_2	0.886	0.913	0.695
V_3	0.918	0.932	0.756
V_4	0.908	0.945	0.745
V_5	0.896	0.715	0.428
V_6	0.578	0.267	-0.250
V_7	0.868	0.863	0.600
B_1	0.878	0.878	0.555
B_2	0.952	0.751	0.547
B_3	0.954	0.913	0.759
B_4	0.873	0.893	0.569
B_5	0.957	0.767	0.584
B_6	0.887	0.888	0.623
B_7	0.885	0.873	0.559
B_8	0.897	0.880	0.597
B_9	0.881	0.872	0.549

Table 6.15: Selected features for each method, including local and global features.

Method	Efron	BHVI
CFS	V_1^G, V_7^G, V_1^I, I_5^I	V_1^G, V_7^G, I_5^I, I_1^C
Relief	C_v^G, A_v^G, V_1^G, P_v^G, V_6^G, A_v^I, V_1^I, P_v^I	C_v^G, A_v^G, V_1^G, P_v^G, V_6^G, V_1^I
M5	V_1^G, V_7^G, I_5^I	V_1^G, V_7^G
SMOReg	V_1^G, V_7^G, V_5^I	V_1^G, V_7^G
SVR-RFE	C_v^G, I_4^G, W_v^G, V_5^I, W_v^I, B_8^C	C_v^G, I_4^G, W_v^G, V_5^I, W_v^I, V_7^C
Union	C_v^G, A_v^G, V_1^G, I_4^G, P_v^G, V_6^G, V_7^G, W_v^G, A_v^I, V_1^I, P_v^I, V_5^I, I_5^I, W_v^I, B_8^C	C_v^G, A_v^G, V_1^G, I_4^G, P_v^G, V_6^G, V_7^G, W_v^G, V_1^I, V_5^I, I_5^I, W_v^I, I_1^C, V_7^C

6.6 Extension to other dataset

In previous sections, both the definition and the analysis of image features were conducted. In order to ensure the robustness of the methodology, a study was conducted to observe how the proposed techniques behave when applied to a different data set. To that end, in this section the IMG_1 dataset is used to validate the proposed approach.

The pairwise correlation was computed also for the features computed on the IMG_1 dataset. The results are depicted in Fig. 6.6. By following the same criteria used with the VID_1 dataset, the features are grouped as depicted in Table 6.16. The same table shows the comparison between both datasets. Each column of the table represents the features that belong to the same group in IMG_1 dataset, and each row, the features that belong to the same group in VID_1 dataset. For example, features A_v, P_v, V_4, B_1, B_4, B_7 and B_8 are grouped together in the VID_1 dataset. However, they belong to two separated groups in IMG_1.

Table 6.16: Groups of features in VID_1 and IMG_1 dataset. Each row represents a group in VID_1 dataset (G_n^V), while each column represents a group in IMG_1 dataset (G_n^I).

		IMG_1								
		G_1^I	G_2^I	G_3^I	G_4^I	G_5^I	G_6^I	G_7^I	G_8^I	G_9^I
VID_1	G_1^V	C_v								
	G_2^V		A_v P_v	V_4 B_1 B_4 B_7 B_8						
	G_3^V		V_1	B_9						
	G_4^V				W_v					
	G_5^V					I_1				
	G_6^V						I_2 I_3 I_5 V_2 V_3 V_7 B_2 B_3 B_6	I_4 B_5 B_2	V_5 B_2	
	G_7^V									V_6

It can be observed how the groups are similar to the ones obtained with VID_1 dataset. The most interesting difference is that features A_v and V_4 are no longer related. Features V_5 and B_2 are alone in a group, as only correlations above 0.7 (absolute value) are taken into account. However, a certain correlation can be noticed with the other features that were grouped together in the previous data set.

Regarding the selected features, Table 6.17 shows the different features chosen by each method in each subset in the Efron scale when taking into account 25 features. The features that are chosen in both data sets are highlighted in bold. The colour in the vessels is more relevant in the IMG_1 data set, as feature V_1 and at least other from V_1, V_2 and V_7 are selected in all but one methods. The average width of the vessels, W_v, also appears more frequently in IMG_1. Besides, all the feature selection techniques obtain larger feature subsets with IMG_1 data set. This means that more information is necessary to evaluate IMG_1, and that the vessels gain relevance, probably because the images are closer to the camera and the hyperaemia level is less variable in this dataset than in VID_1.

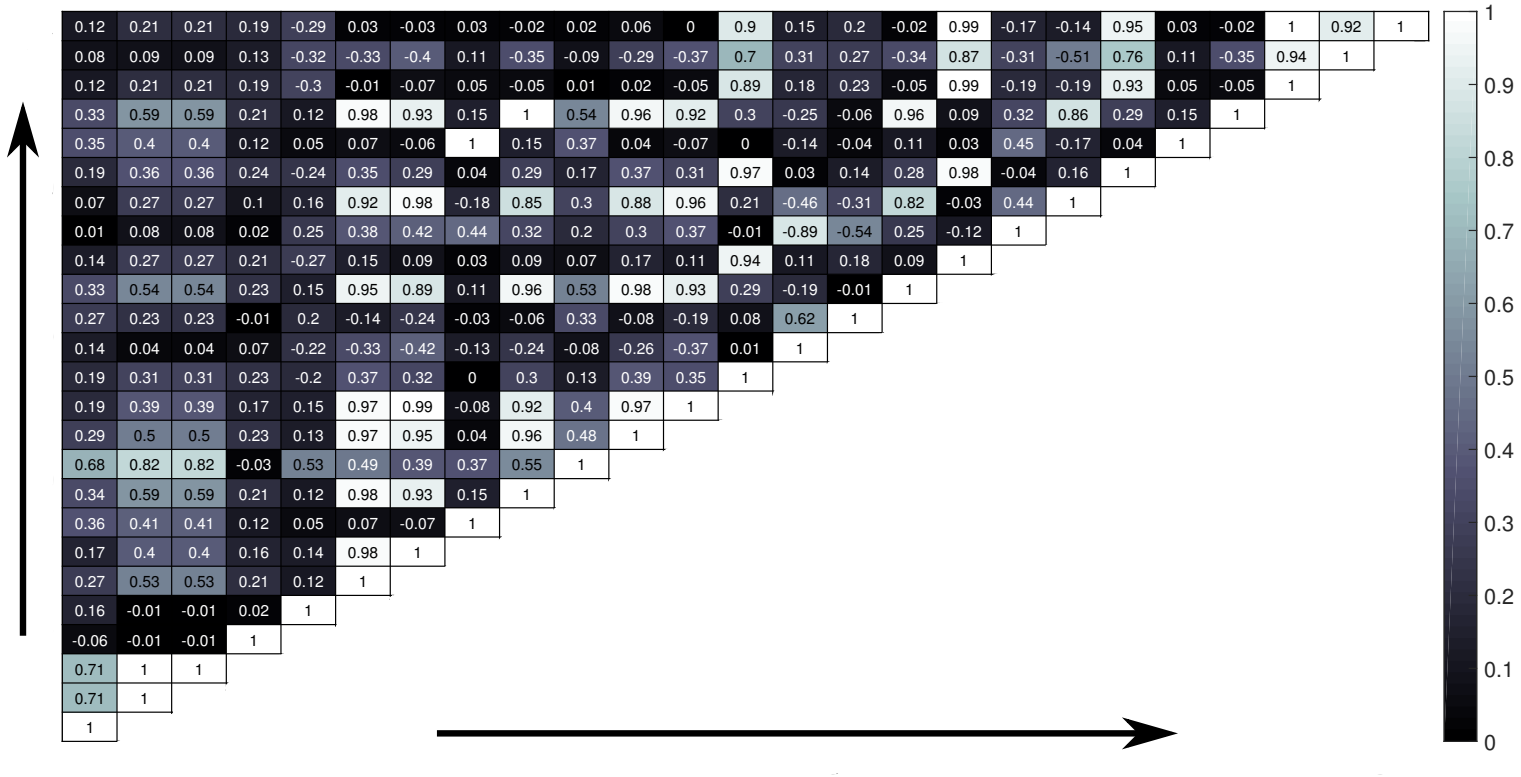


Figure 6.6: Pairwise feature correlation for IMG_1 dataset. Both axis represent the 25 features in the order that they were defined, placed from bottom to top and from left to right.

Table 6.17: Comparison of global features that appear in at least 7 out of 10 folds in VID_1 and IMG_1 datasets. The features that are selected by a method in both datasets are highlighted in bold.

	VID_1		IMG_1	
Method	#	features	#	features
CFS	5	I_1, $\boldsymbol{I_5}$,$\boldsymbol{B_6}$,B_8,$\boldsymbol{W_v}$	6	V_1, I_2, $\boldsymbol{I_5}$, V_7, $\boldsymbol{B_6}$, $\boldsymbol{W_v}$
Relief	4	V_6,$\boldsymbol{I_5}$,$\boldsymbol{V_7}$,$\boldsymbol{B_6}$	8	A_v,V_1,V_2,I_2,P_v,$\boldsymbol{I_5}$,$\boldsymbol{V_7}$,$\boldsymbol{B_6}$
M5	1	$\boldsymbol{B_6}$	5	V_1,V_4,I_5,$\boldsymbol{B_6}$,W_v
SMOReg	3	I_1,$\boldsymbol{I_5}$,$\boldsymbol{B_6}$	5	V_1,V_4,$\boldsymbol{I_5}$,$\boldsymbol{B_6}$,W_v
SVR-RFE	2	A_v, V_5	4	C_v,I_4,P_v,B_3

Table 6.18 depicts the analogue results for the 75 feature vectors. The exact coincidences in feature section have been highlighted in blue, while the features that are selected in both data sets but in a different area appear in green. Most of these selected features that are computed in different areas have a high correlation, as can be observed in Table 6.14, with the exception of I_5^I and I_5^C, that are only mildly correlated. An interesting result is that the relevance of the hue of the vessels is now higher for the VID_1 data set, although feature W_v is still preferred in the IMG_1 data set. Moreover, several of the approaches obtain smaller subsets in the IMG_1 set. Specially interesting is the result obtained by SMOReg, as it consists only of two features, and both are global features. This happens because adding local features causes, with this particular technique, a higher variability among folds. As a consequence, some of the features that were selected in the Table 6.17 are no longer selected, but the local ones that are favoured instead do not appear in enough folds to be considered.

Table 6.18: Comparison of local and global features that appear in at least 7 out of 10 folds in VID_1 and IMG_1 datasets. The features that are selected by a method in both datasets are highlighted in blue or green if they are computed in the same or different areas, respectively.

Method	#	VID_1 features	#	IMG_1 features
CFS	4	V_1^G, V_7^G, V_1^I, I_5^I	8	V_1^G, I_5^G,B_6^G,A_v^I, I_5^C, V_7^C,B_8^C,W_v^C
Relief	8	C_v^G, A_v^G, V_1^G, P_v^G, V_6^G, A_v^I, V_1^I, P_v^I	7	A_v^G, P_v^G,V_7^G,I_2^C, I_5^C,V_7^C,B_6^C
M5	3	V_1^G, V_7^G, I_5^I	4	V_1^G,V_4^G, I_5^G,W_v^G
SMOReg	3	V_1^G, V_7^G, V_5^I	2	V_1^G,I_5^G
SVR-RFE	6	C_v^G, I_4^G, W_v^G, V_5^I, W_v^I, B_8^C	4	I_4^G, W_v^G, W_v^I,A_v^C

6.7 Conclusions

In this chapter, the image features that are used to assess the hyperaemia level were detailed. Then, their relationship with the experts' evaluations as well as with other features was studied. It was found out that is not possible to establish a direct correspondence between the automatic features and the experts' values. Moreover, the relations among features are not always apparent, and some of the features that have a similar formulation have a weak correlation.

Therefore, feature selection methods were applied in order to ensure that the set of employed features have most of the information but does not include redundancy. Five feature selection methods were evaluated in both datasets. The preferred features vary depending on both the grading scale and the dataset, and few features, such as I_5, are selected almost unanimously.

Next, the influence of the region of computation on the relevance of the features was also analysed. When using local features, vessels gain importance, while the background of the conjunctiva is not as relevant as in the whole image. The features selected by the methods are also different than the ones obtained when taking into account only the global values.

Finally, the same five feature selection techniques were applied to a second dataset. The results show that there are some noticeable differences between datasets, which was expected since the characteristics of the images are different. However, some features, such as I_5 (L*a*b* a-channel of the image) or B_6 (red in background in L*a*b*) in the global-only feature experiment, or V_1 (relative vessel-redness) in the local and global features test, are consistently chosen by several methods in both datasets. As it was expected, the red level is the characteristic that stands more consistently. Regarding colourspaces, L*a*b* representation of colours is closer to the human perception and, therefore, its features are closer to the experts' evaluation. Finally, the fact that these features appear in both tests implies that there are some main image features that are relevant despite the variability of the samples and, therefore, that using these features may produce accurate results with new data sets, potentially improving the generalisation capabilities of the methodology.

Chapter 7

From the image features to the grading scale

As Chapter 6 established, the relationship between a single feature and the experts' evaluation is not straightforward. Therefore, more complex approaches are needed, such as combining features in order to ensure that there is enough information to represent the problem. This task has a strong relevance in the process, as it will decide the contribution of each feature to the final result.

However, combining features creates a new obstacle to overcome, as each feature has a different range of values, apparently unrelated to the grading scales. Therefore, the last step of the methodology is to map the feature values to the evaluations in a given scale. Since this is a complex transformation, machine learning techniques were used.

7.1 Machine learning techniques

There are several options available to establish complex relationships in a dataset. As it was mentioned in Chapter 1, the grading scales used to assess hyperaemia are collections of a small number of prototypes. As there are several intermediate cases between each pair of prototypes, there is a certain subjectivity in the manner that

experts use these scales. For example, the optometrists evaluated IMG dataset[1] in 0.25 intervals. However, the VID dataset[2] was evaluated in 0.1 intervals. Therefore, given the nature of the data, two main approaches can be tackled:

Regression methods. Since the gradings can be seen as an array of continuous values, regression methods can be applied.

Classifiers. Usually specialists are not confident enough to provide an evaluation more precise than a certain threshold, where the differences between consecutive categories are too small to be noticed by the human eye. Therefore, the problem can be tackled as a common classification problem.

At first glance, it may seem that classifiers offer a closer representation to the problem, as some regression approaches require assumptions that the data follows a certain distribution or has a given structure. However, regression approaches have the advantage of being able to generalise intermediate classes that are not well represented in the dataset. As it was explained in Chapter 2, some of the classes, specially the extreme ones, are uncommon and, therefore, there are not available samples on the category. Furthermore, classification methods have an additional step that must be faced: the number of classes in the experts' evaluation is too large in comparison with the number of images of the dataset. Therefore, the dataset must be split in order to create larger groups of images within a reasonable range of values.

7.1.1 Regression approaches

Some of the regression approaches needed additional steps in order to transform the inputs and outputs, as the algorithms are used more commonly in classification problems. The following regression approaches were tested:

Decision trees (DT). Decision trees [82] are structures that establish tree-like structures by creating rules in order to divide the inputs. The implementation used in this work uses the CART algorithm [83].

[1]Images captured by the School of Optometry and Vision Sciences (Cardiff University)

[2]Videos captured by the Optometry Group (University of Santiago de Compostela)

K-nearest neighbours (KNN). The instance based methods [84] do not perform an explicit generalisation in the manner that other methods do. Instead, they compare each new instance that appears with the instances that appeared during the training stage. K-nearest neighbours [85] is an algorithm that takes into account the value of the current sample, and the values of the k closest neighbours. This technique can be used in both regression or classification scenarios.

Learning vector quantization (LVQ). The learning vector quantization [86] is within the group of the artificial neural networks (ANNs). It consists of a competitive and a linear layer. The first one consists of a series of neurons where one is chosen as the most appropriate to represent the input. The second one transforms the output of the first layer into target classes, previously defined by the user.

Multi-layer perceptron (MLP). The multi-layer perceptron [87] is one of the most widely used methods for finding complex relationships in data with similar problems. It is an ANN that consists of several layers. It has a feed-forward structure, as each node of each layer is fully connected with the next layer. Each neuron in the middle (hidden) layers will apply a non-linear activation function. The training is performed with a backpropagation algorithm.

Naive-Bayes (NB). The naive Bayes approach [88] denotes a family of classifiers based on the application of Bayes' theorem. They are probabilistic classifiers that make strong independence assumptions between the analysed features, the reason why they are called *naive*.

Partial least squares (PLS). The partial least squares approach [89] is a regression technique that creates a linear regression model by projecting the predicted variables to a new space.

Radial basis function network (RBFN). The radial-basis function network [90] is also a feedforward ANN, with the particularity of using radial basis functions as activation functions. It was selected because it operates in a similar fashion to the experts when they measure hyperaemia. The network weighs the closeness from

the input to each prototype, while the optometrist usually assigns the evaluation depending on the closeness to each grade of the scale.

Random forest (RF). The random forest [91] is an ensemble method that creates and evaluates several decision trees, outputting their mean prediction. It is also widely used and suited to this kind of problem, as ensemble methods are expected to provide better results than single decision trees.

Self-organising map (SOM). The self-organising map [92] belongs also to the ANN group of techniques. It applies competitive learning, and takes into account neighbourhood relationships established depending on the topology.

Support vector regression (SVR). The support vector regression [93] is a variant of the support vector machine (SVM) applied to regression problems.

Model tree with M5 algorithm (M5P). It is the same implementation of the model tree algorithm that was applied in feature selection [94, 79] in Chapter 6.

Linear regression (LR). The linear regression models the relationship between a dependent and a explanatory variable [95].

These methods were selected in order to cover a wide spectrum of machine learning techniques. The neural networks are good at finding complex relationships in datasets, but their results can be difficult to explain, which can be a drawback when the objective is to understand the underlying relationships of the data. Moreover, the training process is generally slow, specially when compared to other methods, such as decision trees. However, as the objective in the problem at hand is to train the network once and then perform only predictions, this slowness should not be a hindrance.

7.1.2 Classifiers

Once a data splitting was decided on the experts' evaluations, the classifier approaches can be applied. The following classifiers were tested:

Bayes network (BN). The bayes network [96, 97] consists in a directed acyclic graph that represents a set of random variables, as well as their dependencies. It is

a probabilistic model, commonly used to represent probabilistic relationships among variables.

Decision table (DT). The decision table [98] models rule sets in a compact manner. As other rule-modelling options, it associates each rule with its corresponding action.

K-nearest neighbours (KNN). An algorithm of the instance based family applied to classification [84]. The implementation is the same that was used for regression, k-nearest neighbours.

Decision tree J48 (J48). A decision tree that follows the implementation C4.5 [99].

Naive bayes (NB). The naive Bayes approach [88] is the same defined in the regression approaches.

One rule (OR). The one rule [100] is an algorithm that consists of two main steps. First, it produces a rule for each predictor in the data. Next, it selects the rule that achieves the lowest total error. Each rule is based on the frequency table of each predictor against the target. This algorithm is usually less accurate than most state-of-art methods, but its results are easy to understand.

Random forest (RF). The random forest [101] is the same approach proposed in the regression techniques.

Support vector machine (SVM, SMO). The classification version of the support vector machine [102], equivalent to the SVR approach for regression. Additionally, a version with sequential minimal optimisation (SMO) was used [103, 104, 105]. This second approach, also employed by one of the feature selection techniques (Chapter 6), is one of the available options for training a SVM.

These algorithms were also chosen trying to cover most types of classification methods. Each of the proposed techniques has its own advantages and disadvantages, such as computational complexity, difficulty of the parameter tune up, or generalisation capability.

In order to apply a classification approach to the data, the evaluations must be grouped to form discrete classes. In order to create these groups, the following options were considered:

Using integer and half-integer values. This idea is supported by the works that conclude that experts usually grade taking these characteristic values as a reference [15].

Using one decimal. This is the interval size of VID dataset, and generally the maximum precision of human experts when grading.

Using integer, half and quarter values. This is the interval size of IMG dataset. Incidentally, it provides a middle point between the other two approaches and, therefore, it helps with the analysis of how the values evolve with the gap between classes.

7.2 Validation procedure

For the validation with the VID_1 and VID_2 datasets, the ground truth for the systems is the average of four gradings: the two evaluations performed by E_1 and the two evaluations performed by E_2.

For the validation with the IMG_1 dataset, the ground truth for the systems is the average of two values, the evaluations performed by the experts with the highest correlation between them (E_2 and E_3).

Once the output of the machine learning technique is obtained, this value is compared with the ground truth of that image by means of the mean squared error (MSE). The mean squared error is an statistic defined by the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{7.1}$$

where n is the number of elements, $\hat{Y}$ is the vector comprising the estimated values (automatic outputs) and Y is the vector of the expected values (manual evaluations).

The MSE is commonly used as a means to compare regression techniques. It is always positive, and the lower, the better. There are other usual parameters that can be used for the same purpose, such as the coefficient of determination (R^2) or the mean absolute error (MAE), although the obtained values in this work are related for the three estimators.

As a 10-fold cross-validation approach is used during the whole process, the MSE depicted in the results is the validation error, that is, the error is computed from the output of the test set in each fold. All the experiments were run with 100 iterations of cross-validation.

When obtaining the output of a regression system, the cross-validation error was computed. However, the procedure applied to compute the MSE with the classifiers is slightly different, as the output is not continuous. The success rate (SR) was computed for the classifiers, by comparing the systems' outputs with the expected values. In order to ensure an objective comparison, the same success rate was also computed with the regression approaches. To that end, an instance is counted as correctly classified if the expected class is the closest to the given output.

7.3 Regression results

As a previous step, the best configuration for each system was empirically determined. To that end, the average 10-fold cross-validation error was used as goodness measure. The parameters that achieved the best results are depicted in Tables 7.1 and 7.2.

Table 7.1: Parameters of the regression methods.

Method	Parameters
DT	minimum leaf size = 3
	minimum parent size = 6
KNN	number of neighbours = 1
	distance = cosine
LVQ	number of neighbours for the classification stage = 3,
	dimensions = 8 (Efron) and 6 (BHVI)
MLP	layer configuration = [40 16]
	activation function = hyperbolic tangent sigmoid
	training function = Bayesian regularization backpropagation
	based on Levenberg-Marquardt optimization
	epochs = 1000
	weight initialisation = Nguyen-Widrow
NB	determination of prototypes using a KNN model with number of neighbours = 3
	distribution type = mvmn (Efron) and kernel with normal kernel (BHVI)
PLS	number of components = min(number of features, 8)
RBFN	spread = 0.4
	error goal = 0.03
RF	number of trees = 60 (Efron) and 40 (BHVI)
	minimum leaf size = 10
SOM	competitive layer size = 8 (Efron) and 6 (BHVI)
	number of neighbors = 3
	topology function = one-dimensional random pattern
	distance function = Manhattan
	configuration of the MLP = [10]
SVR	type = ν-SVR (Efron) and ϵ-SVR (BHVI)
	kernel = sigmoid (Efron) and radial basis function (BHVI)
	$\gamma = 2^{-12}$ (Efron) and 2^{-10} (BHVI)
	C = 2^8 (Efron) and 2^4 (BHVI)
M5P	minimum instances per leaf = 4

The correlation and kappa results established in Chapter 6 that none of the individual features has a straightforward relationship with the experts evaluations. Moreover, feature selection approaches produced sets consisting of more than one features in all but one tests. Therefore, individual features will not produce optimal results. However, in order to establish a numeric comparison, each of the individual features was used to train and test three of the regression techniques. To that end, the MLP, RF and RBFN were selected, as the three of them are used in similar environments in the state

Table 7.2: Parameters of the classifiers.

Method	Parameters
BN	search algorithm = K2 [106]
	Alpha value for the estimator = 0.5
SVM	type = C-SVC
	kernel = radial basis function
SMO	Pearson universal kernel ($\omega = 1.0$, $\sigma = 1.0$)
KNN_1	neighbours = 1
KNN_3	neighbours = 3
DT	evaluation measure = RMSE
J48	confidence factor for pruning = 0.25
	minimum instances per leaf = 2
RF	number of trees = 100

of the art and, thus, they are expected to provide good results. Table 7.3 depicts the obtained values. The VID_1 dataset was used to compute the MSE of the algorithms trained with the individual features.

The MLP is the approach that obtains the lowest MSE in most of the cases for the BHVI scale. Yet, the results are inconsistent, as the average error is lower for the RF approach. Also, the RF outperforms the other approaches in the Efron scale, as it obtains the lowest individual MSE, and the lowest average. In general, the regression methods seem to offer a closer representation to the BHVI scale than to the Efron, as errors about 0.1 are obtained with several features.

Table 7.3: MSE values for three regression techniques applied to single features in the Efron and BHVI scales. The best value for each feature is highlighted.

Feature	Efron			BHVI		
	MLP	RBFN	RF	MLP	RBFN	RF
C_v	0.301	0.220	0.204	0.095	0.136	0.125
A_v	0.454	0.227	0.225	0.106	0.146	0.145
P_v	0.569	0.227	0.224	0.109	0.146	0.147
W_v	0.315	0.240	0.246	0.128	0.147	0.162
I_1	0.099	0.222	0.207	0.060	0.143	0.137
I_2	0.054	0.223	0.152	0.037	0.142	0.095
I_3	0.360	0.224	0.162	0.039	0.141	0.107
I_4	0.313	0.225	0.150	0.052	0.141	0.094
I_5	0.260	0.214	0.134	0.094	0.139	0.086
V_1	0.082	0.216	0.222	0.181	0.140	0.148
V_2	0.402	0.218	0.152	0.261	0.138	0.096
V_3	0.128	0.222	0.197	0.115	0.143	0.130
V_4	2.310	0.226	0.228	0.925	0.143	0.145
V_5	0.142	0.202	0.156	0.172	0.133	0.103
V_6	0.135	0.223	0.250	0.115	0.143	0.155
V_7	0.314	0.196	0.149	0.213	0.131	0.099
B_1	0.196	0.223	0.204	0.137	0.143	0.127
B_2	0.185	0.223	0.147	0.081	0.144	0.095
B_3	0.078	0.223	0.226	0.291	0.142	0.148
B_4	0.355	0.230	0.236	0.392	0.145	0.150
B_5	0.175	0.223	0.146	0.089	0.141	0.095
B_6	0.611	0.215	0.135	0.198	0.137	0.085
B_7	0.585	0.240	0.198	0.209	0.141	0.124
B_8	0.270	0.226	0.210	0.743	0.142	0.138
B_9	0.387	0.228	0.181	0.183	0.146	0.112
Mean	0.3632	0.2222	0.1896	0.2010	0.1413	0.1219

Table 7.4 depicts how the MSE is reduced with the combination of features in comparison to the individual results. Moreover, the error values for several feature selection subsets in some approaches are lower than with the full set of features. Therefore, the feature selection techniques not only reduce the complexity of the problem, but also are able to improve the results. The approach that achieves the best results in both scales is the PLS. The RF also obtains good results in both scales, and so it does the MLP, specially in the BHVI scale. The SMOReg subset for the BHVI scale favours also the SVR and SOM approaches, and the whole set of features, the DT and KNN methods. Finally, LVQ or NB obtain the worst results, implying that the methods are not suited to the problem. The values are consistently better for the BHVI scale.

Table 7.4: Comparison of the MSE values of each regression technique and feature combination (global-only features). The lowest MSE for each regression technique is highlighted.

Efron scale							
Method	All features	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.119	0.173	0.158	0.184	0.164	0.200	0.127
KNN	0.109	0.231	0.199	0.225	0.203	0.244	0.119
LVQ	0.427	0.449	0.467	0.383	0.289	0.332	0.482
MLP	0.181	0.205	0.111	0.234	0.098	0.118	0.166
NB	0.700	0.742	0.725	0.692	0.742	0.750	0.750
PLS	0.048	0.063	0.088	0.117	0.079	0.117	0.050
RBFN	0.380	0.380	0.380	0.380	0.380	0.361	0.380
RF	0.067	0.111	0.108	0.131	0.114	0.128	0.079
SOM	0.210	0.210	0.213	0.127	0.210	0.157	0.209
SVR	0.228	0.228	0.228	0.118	0.228	0.228	0.228
BHVI scale							
Method	All features	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.091	0.141	0.119	0.111	0.138	0.159	0.102
KNN	0.085	0.174	0.102	0.147	0.128	0.195	0.096
LVQ	0.206	0.245	0.228	0.203	0.234	0.175	0.238
MLP	0.146	0.126	0.123	0.074	0.087	0.088	0.139
NB	0.181	0.153	0.156	0.265	0.158	0.204	0.161
PLS	0.041	0.050	0.062	0.069	0.070	0.089	0.042
RBFN	0.248	0.248	0.248	0.248	0.554	0.245	0.248
RF	0.052	0.089	0.066	0.094	0.100	0.100	0.060
SOM	0.157	0.157	0.113	0.153	0.097	0.118	0.157
SVR	0.172	0.172	0.160	0.177	0.087	0.196	0.172

Therefore, it can be observed that feature selection methods are able to successfully reduce the number of features used in the computation of the hyperaemia level while maintaining the MSE values that are obtained by using all the features. Furthermore, the benefits are not only related to discovering the importance of a given feature, but are also reflected in a significant reduction of the computation time, by 60% and 40% for the Efron and BHVI scales, respectively.

7.4 Regression vs classification

The continuous output of the optometrists was divided in classes. Several configurations were tested, by varying the step between classes (0.5, 0.25 and 0.1). The obtained results are depicted in Table 7.5.

Table 7.5: Classification results for steps 0.5, 0.25 and 0.1. The best SR and lowest MSE for each step and gradings scale are highlighted.

		step=0.5				step=0.25				step=0.1			
		Efron		BHVI		Efron		BHVI		Efron		BHVI	
		SR	MSE	SR	MSE	SR	MSE	SR	MSE	SR	MSE	SR	MSE
Classification	BN	46.7	0.094	53.3	0.106	20.0	0.056	32.4	0.073	9.5	0.023	11.4	0.030
	NB	38.1	0.126	43.8	0.141	21.9	0.080	31.4	0.089	9.5	0.040	14.3	0.049
	SVM	41.9	0.129	44.8	0.158	21.0	0.093	25.7	0.114	10.5	0.044	12.4	0.057
	SMO	48.6	0.082	56.2	0.096	21.0	0.051	36.2	0.064	7.6	0.030	11.4	0.023
	KNN_1	40.0	0.133	51.4	0.139	27.6	0.085	31.4	0.106	10.5	0.044	20.0	0.052
	KNN_3	41.0	0.088	44.8	0.103	15.2	0.060	24.8	0.072	9.5	0.028	13.3	0.034
	DT	46.7	0.073	54.3	0.086	21.0	0.050	33.3	0.061	9.5	0.023	10.5	0.030
	OR	35.2	0.144	45.7	0.155	24.8	0.089	25.7	0.114	13.3	0.042	11.4	0.057
	J48	39.0	0.125	53.3	0.116	28.6	0.077	31.4	0.093	8.6	0.037	9.5	0.048
	RF	41.0	0.076	58.1	0.080	18.1	0.051	35.2	0.061	6.7	0.025	17.1	0.030
Regression	KNN	46.7	0.177	56.2	0.125	25.7	0.148	39.1	0.091	11.4	0.138	14.3	0.086
	LR	54.3	0.163	55.2	0.121	34.3	0.141	28.7	0.089	16.2	0.139	15.2	0.081
	M5P	47.6	0.144	58.1	0.104	28.6	0.142	34.3	0.088	18.1	0.117	15.2	0.075
	MLP	36.2	0.238	50.5	0.178	31.4	0.200	35.2	0.120	9.5	0.176	14.3	0.106
	RBFN	40.0	0.219	49.5	0.145	20.0	0.196	25.7	0.131	8.6	0.194	4.8	0.126
	SVR	41.9	0.245	44.8	0.160	21.0	0.223	25.7	0.145	7.6	0.219	7.6	0.141

The error values grow in direct proportion to both the step and the number of correctly classified images. This was the expected behaviour, as a misclassified output

generates a worse error if the gap between values is wider. The results for the SMO classifier are good with all approaches. Some of the approaches, such as RF with step 0.25, obtain a perfect classification of the test set. However, the main goal of the experiment is to obtain a MSE as low as possible, as the number of correctly and incorrectly classified images can be misleading, specially with the broader gaps. Thus, an overfitted model will achieve good results regarding success rate, but at the cost of a higher MSE.

If both types of methods are compared, it can be observed how the regression methods perform generally better regarding success rate, with the exception of the SMO classifier. However, their error is also higher, as regression methods do not output the exact class value, so most of the predictions add some error. Thus, the error is higher than the best cross-validation error achieved in previous experiments (Table 7.4).

In order to better illustrate this issue, a final test was run, in which the number of images that are classified in contiguous classes with step 0.1 is analysed, since this scenario will provide an accurate classification too. Figures 7.1 and 7.2 depict how the success rate varies when taking into account the instances classified in neighbouring classes, with the x-axis showing the tolerance margins and the y-axis showing the percentage of correct classifications for that given tolerance.

There are certain differences regarding the chosen scale. For example, in the BHVI scale the methods are able to correctly classify more instances with lower margin levels. However, in the Efron scale there is only one method that achieves a 90% of success rate with the maximum margin. Nevertheless, the approaches that achieve the best results are the regression techniques for both scales. The system is able to classify correctly 90% of the instances in the Efron scale with a ± 0.5 margin. The results improve in the BHVI scale, as a ± 0.4 margin is enough to achieve more than a 90% of success rate. This is probably caused by the nature of the prototypes of each scale, as the BHVI scale consists of real eye photographs while the Efron scale is a collection of drawings. Moreover, the distribution of the values is also different, as the steps between photographs in the BHVI scale vary while the steps in the Efron scale are more evenly

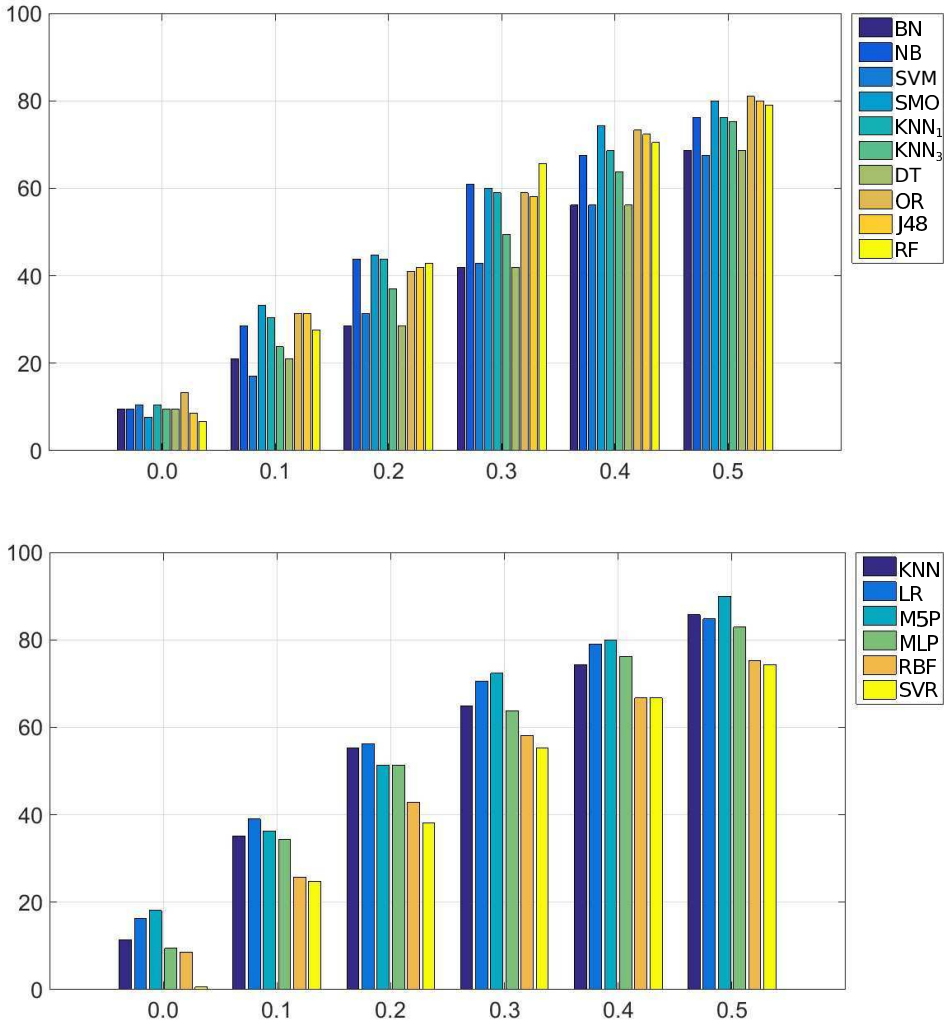


Figure 7.1: Evolution of the success rate in the Efron scale. x-axis represents the margin and y-axis, the success rate. Top: classification techniques. Bottom: regression techniques.

spaced. Therefore, the experts' evaluations change as a direct consequence of the scale and these differences are reflected in the automatic outputs.

As a conclusion, even though both groups of methods can be valid to solve the problem at hand, results show that the automatic methodology benefits more from regression approaches, as the results are better for these methods in both grading scales.

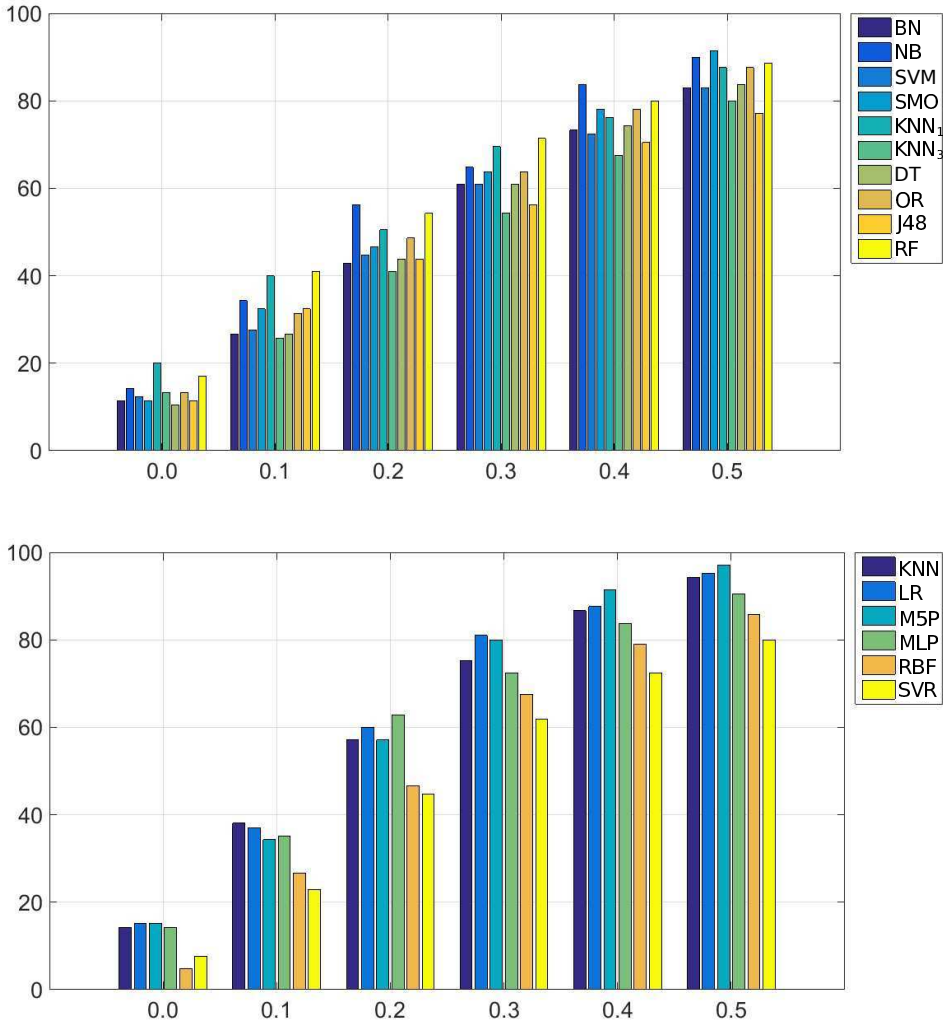


Figure 7.2: Evolution of the success rate in the BHVI scale. x-axis represents the margin and y-axis, the success rate. Top: classification techniques. Bottom: regression techniques.

7.5 Local vs global features

Since the regression techniques provided better results than the classifiers, the experiments regarding the influence of local features in the results were performed with the regression approaches. Tables 7.6 and 7.7 depict the MSE values obtained when applying the machine learning techniques to the feature selection sets obtained from the local and global feature set.

Table 7.6: MSE combination of the local and global features for all systems in VID_2 dataset (Efron scale). The lowest MSE for each technique is highlighted.

Method	All	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.111	0.103	0.196	0.099	0.121	0.215	0.112
KNN	0.130	0.129	0.225	0.160	0.147	0.254	0.138
LVQ	0.459	0.518	0.314	0.494	0.456	0.426	0.426
MLP	0.188	0.062	0.192	0.060	0.080	0.143	0.173
NB	0.717	0.642	0.633	0.658	0.592	0.642	0.575
PLS	0.073	0.058	0.181	0.058	0.071	0.127	0.065
RBFN	0.380	0.380	0.380	0.380	0.380	0.372	0.380
RF	0.066	0.071	0.125	0.070	0.076	0.132	0.068
SOM	0.206	0.124	0.124	0.122	0.119	0.153	0.124
SVR	0.228	0.228	0.228	0.228	0.228	0.251	0.228

Table 7.7: MSE combination of the local and global features for all systems in VID_2 dataset (BHVI scale). The lowest MSE for each technique is highlighted.

Method	All	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.090	0.077	0.154	0.089	0.089	0.167	0.088
KNN	0.096	0.098	0.183	0.124	0.124	0.178	0.104
LVQ	0.245	0.214	0.202	0.246	0.246	0.203	0.226
MLP	0.143	0.054	0.135	0.055	0.055	0.118	0.127
NB	0.141	0.169	0.227	0.173	0.173	0.255	0.199
PLS	0.058	0.046	0.088	0.055	0.055	0.116	0.050
RBFN	0.248	0.248	0.248	0.248	0.248	0.252	0.248
RF	0.053	0.053	0.095	0.058	0.058	0.118	0.052
SOM	0.156	0.173	0.088	0.089	0.089	0.117	0.088
SVR	0.172	0.172	0.171	0.176	0.176	0.105	0.172

For the Efron scale, the best values are achieved with PLS with the CFS or M5 feature set. The RF also obtains good results combined with several feature selection methods, as well as with the whole feature set. The best value for BHVI scale is obtained with the PLS in the CFS feature set. BHVI obtains better results than Efron in general, with the RF and the MLP approaches obtained low MSE in several cases, and other approaches obtaining ocassionally MSE below 0.1. As the usual step of gradings is 0.5, this value is the maximum acceptable difference between the expert and the algorithm output. Given that the MSE is the square of the differences, an

MSE below 0.25 is suitable. There are many approaches that fulfil this rule, so it can be concluded that the methodology is able to mimic a human expert in both scales.

Finally, if these results are compared with the results of Table 7.4, it can be observed how the MSE is maintained or even reduced in both scales for a large number of configurations, hence confirming that the consideration of local features can improve the overall performance of the system. This improvement is higher in the Efron scale.

7.6 Extension to other datasets

In order to observe the behaviour of the methodology in a real-world environment, a validation was performed with a different dataset. Table 7.8 shows the results obtained using the feature sets computed in the IMG_1 dataset with the proposed algorithms. The experiments were performed with global features as well as local and global features combined.

Regarding the results for the global features, the best method is PLS in all the cases, followed by the MLP or RF approaches. These results are similar to the ones obtained in VID_2 dataset. Observing the feature selection methods, the best results are obtained with both wrappers, M5 and SMOReg.

The results obtained with global and local features are similar. The best regression techniques are the same three algorithms, PLS, RF and MLP. The best result is achieved with PLS and the M5 feature selection technique.

It must be noted that, while the lowest error with VID_2 dataset was obtained with all the features for the global-only experiment, the tests with IMG_1 dataset show better results with reduced subsets in both scenarios.

Table 7.8: Comparison of MSE values for features appearing in 7 out of 10 folds in the IMG_1 dataset.

25 features							
Method	All features	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.116	0.106	0.109	0.093	0.093	0.124	0.108
KNN	0.097	0.123	0.146	0.111	0.111	0.138	0.117
LVQ	0.253	0.247	0.319	0.247	0.247	0.250	0.259
MLP	0.121	0.080	0.105	0.063	0.063	0.088	0.097
NB	0.355	0.419	0.419	0.390	0.390	0.341	0.383
PLS	0.060	0.058	0.058	0.051	0.051	0.071	0.054
RBFN	0.204	0.204	0.204	0.204	0.204	0.203	0.204
RF	0.070	0.070	0.072	0.067	0.067	0.080	0.069
SOM	0.158	0.110	0.110	0.110	0.110	0.115	0.110
SVR	0.161	0.161	0.161	0.161	0.161	0.162	0.161
75 features							
Method	All features	CFS	Relief	M5	SMOReg	SVR-RFE	Combination
DT	0.116	0.107	0.117	0.092	0.095	0.189	0.109
KNN	0.125	0.104	0.161	0.112	0.165	0.301	0.130
LVQ	0.250	0.256	0.255	0.249	0.216	0.230	0.245
MLP	0.162	0.093	0.102	0.055	0.060	0.142	0.123
NB	0.383	0.411	0.418	0.390	0.404	0.340	0.383
PLS	0.099	0.056	0.067	0.051	0.058	0.122	0.056
RBFN	0.204	0.204	0.204	0.204	0.206	0.295	0.204
RF	0.069	0.064	0.077	0.070	0.072	0.128	0.065
SOM	0.158	0.110	0.088	0.110	0.110	0.149	0.110
SVR	0.161	0.161	0.161	0.161	0.161	0.162	0.161

7.7 Conclusions

This chapter described the last step of the automatic methodology, the transformation of the image features to the final output, as that output must have a direct correspondence to the grading scale that is being used.

As grading scales are collections of discrete prototypes, but are commonly handled as a continuous range of values in practice, there are two possible choices for this last step: classification or regression approaches. As there was not enough support to clearly discard one of them, both were implemented and tested. The results favour the regression approaches, as the number of classes is too high in comparison to the available number of samples.

Thus, several regression techniques were evaluated, taking into account the feature sets obtained in the previous chapter, this is, all the local and global features and the feature selection subsets. The results show that the best techniques are PLS, RF and MLP for all the experiments. However, the best feature selection method varies. For the IMG_1 set, both M5 and SMOReg wrappers obtain the lowest MSE with both global-only and global and local features. For the VID_2 dataset, CFS and the union of subsets achieves the best results with global features in both scales, while the complete feature set or the union of subsets give the best results for local and global features.

In view of the results, it can be concluded that the automatic methodology for bulbar hyperaemia grading is able to mimic the experts' behaviour, as the outputs of the machine learning techniques are comparable to the clinicians' evaluations. Moreover, the regression algorithms that obtain the best results are the same in two datasets with different characteristics. Hence, the methodology is general enough to be applied to additional datasets and maintain the accurate results.

Part III

Bringing the methodology to open scenarios

Chapter 8

Repeatability of the methodology

Once the methodology has been implemented and tested, and in view of some of the results, a reasonable concern arises. Is the methodology general enough to ensure that small changes in the images that do not affect the optometrist's evaluation, will also not affect the automatic results? By looking at the characteristics and values obtained by the two data sets, VID and IMG, it can be observed that certain steps of the methodology are more or less adapted to each set, as each data set is obtained under different conditions. The device, the environmental conditions, the illumination and the specialist that obtained the media are different. When analysing real world images and videos, the acquisition process for the inputs is vital, as the changes in the conditions have a strong effect on the outputs, and may hinder or even prevent the application of the methodology.

Therefore, the fact that large changes in the inputs impact similarly the results is not directly solved by the methodology, and has its roots on a non-standarised acquisition of the media. Nevertheless, if this acquisition follows a similar pattern for each image, the results should not vary, even when there is a certain variability on the conditions.

The main objective of this chapter is to prove that the methodology obtains repeatable results under similar conditions. This implies that when an image suffers slight changes that do not alter the grading, such as when two images are taken with and without contact lenses, the methodology remains unaffected as well (Fig. 8.1).

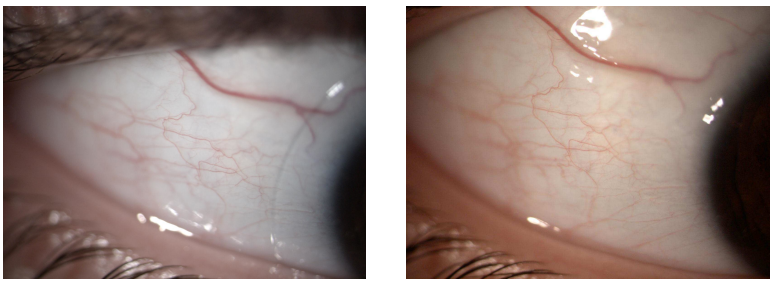


Figure 8.1: Two images of the same eye with and without contact lenses. It can be observed how one of the images was taken under a brighter light, which can affect the colour based features of the image.

8.1 Variations in the images

The IMG data set provides several images from the same eye under different circumstances, so that the complete set is used during this analysis.

The objective is to measure the impact of certain alterations in the methodology. To that end, two image sets were selected. Each set consists of 10 pairs of images. In each pair, one image (*altered*) shows the eye with a given condition or *alteration*, and the other depicts the same eye under optimal conditions to perform the hyperaemia evaluation (*reference*). A grading in the Efron scale was performed for each image. The first image set, labelled S_{blue}, depicts the images with and without remains of a blue dye (Fig. 8.2, top) and the second image set, labelled S_{cont}, shows images with and without contact lenses (Fig. 8.2, bottom).

The images in S_{blue} set belong to the same checkup, which guarantees that they were taken minutes apart. This is the ideal situation for the repeatability study, as it has been noted how hyperaemia can vary through time. Unfortunately, the images in S_{cont} belong to different checkups, but the special characteristics of the data set prevent this being a drawback. It is known that an increase of hyperaemia can be associated to contact lenses wear [107, 108]. However, in the IMG data set the conclusions of the study[1] support the view that the variation in hyperaemia level between the two first checkups was too little to be relevant. This discrepancy is probably caused by the difference in the time exposed to contact lenses wear, as the studies where the contact lenses cause higher hyperaemia refer to a prolonged wearing (8 to 16 hours), while this

[1] http://research.cardiff.ac.uk/converis/portal/Project/2525952

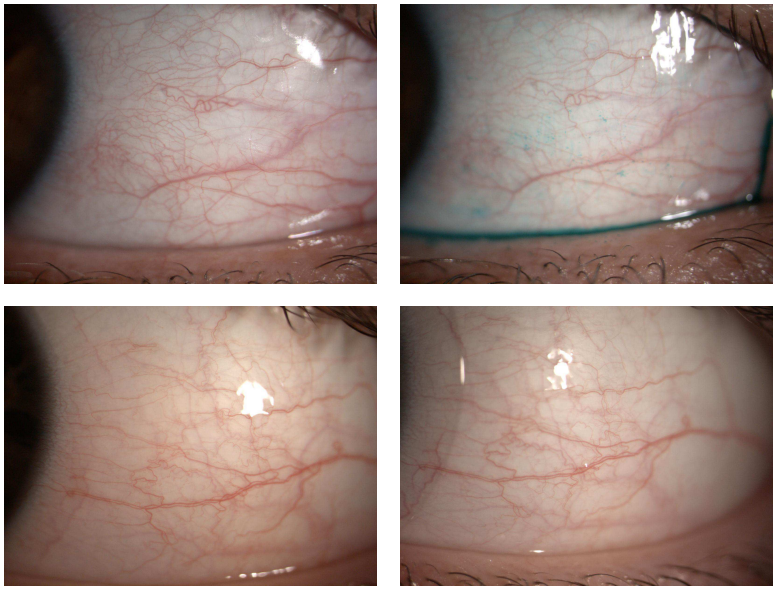


Figure 8.2: A pair of images from each set used during the repeatability study. Top: S_{blue}. Bottom: S_{cont}.

study required a minimum of 4 hours of wearing. Besides, all the patients on the study were healthy, while this condition was not a requisite in other works.

Each stage of the methodology was tested separately in order to prevent bias stacking in the results. First, for the segmentation of the region of interest, the split and merge algorithm was used, as it provides good results in IMG data set. Then, the image features were computed for the conjunctiva. Feature selection techniques were applied to the whole image set. Finally, the image features were mapped to the values in the Efron scale. For the sake of brevity, only three of the techniques that obtained the best results in previous tests were used, these were, the MLP, RF and PLS approaches.

For the validation of the conjunctiva segmentation, the automatic method is applied to each image of the subsets S_{blue} and S_{cont}. Then, the results are compared with the manual segmentation, calculating the sensitivity, specificity, accuracy and precision. As the objective is to compare the results obtained on the pairs of *reference* and *altered* images, the statistical measures will be computed in both cases, and then compared.

In order to measure the effect of the alterations in the feature computation, the 25 image characteristics were computed in both images of each pair. Then, for each

feature, the mean and standard deviation were calculated in both S_{cont} and S_{blue}, making a distinction between *reference* and *altered* images. Next, the coefficient of variation (CV) was computed. The CV is commonly employed in repeatability studies, as it provides insight on the amount of variability relative to the mean of a population.

For the last step of the methodology, both S_{cont} and S_{blue} sets were presented to regression techniques previously trained with the IMG_2 dataset that is made up of the remaining 875 images from the IMG data set, and 100 iterations of 10-fold cross-validation. Each system received each image subset (S_{blue_ref}, S_{blue_alter}, S_{cont_ref}, and S_{cont_alter}) and produced an output. Next, the mean square difference between both outputs was computed as follows:

$$diff_S = avg((output_{S_{ref}} - output_{S_{alter}})^2) \tag{8.1}$$

Finally, four feature selection techniques were applied: the two filter methods, CFS and Relief, and the two wrapper methods, M5 and SMOReg.

8.2 Results

This section presents the results of the experiment of each stage of the process. First, the analysis of the dataset is presented. Next, the segmentation of the conjunctiva is validated. Then, the effect that each alteration of the images has in the features is studied. Finally, the results for the regression systems are explained.

8.2.1 Analysis of the expert's evaluations

In order to establish a gold standard for the subsequent experiments, a variability study was conducted. The evaluation of the optometrist through the four checkups for each patient was analysed. Consecutive checkups are compared pairwise, as their images were taken sequentially in time. Thus, one comparison was performed between checkups 1 and 2 and another, between checkups 3 and 4. That is, between each naked-eye checkup and the consecutive contact lens-wearing one, where the highest variability is expected. For the first pair, the average variations of the expert's evaluations in the

full image set for the right and left eyes are 0.193 and 0.207, respectively. For the second pair, the average variations for the right and left eyes are 0.193 and 0.157, respectively. Half of the images do not vary their evaluation. The complete distribution is depicted in Fig. 8.3. It must be noted that, as described in Chapter 2, the expert that evaluated IMG_2 dataset divided each image in four areas, and labelled each one independently. Thus, the grading of the image is an average of these values. That is why some images have a narrower variability than expected given the granularity of the evaluations.

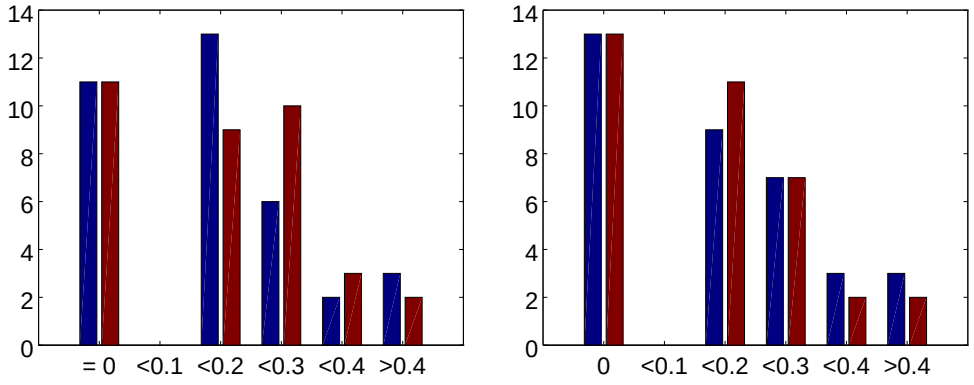


Figure 8.3: Distribution of the variations on the right and left eyes through consecutive checkups. The x-axis depicts the amount of variation and the y-axis, the number of patients that present a range of variation. Left: comparison of C_1 and C_2. Right: comparison of C_3 and C_4.

While the optometrist had the knowledge regarding which side of the eye he/she was looking at, the automatic system cannot access this information. Thus, the variations of the evaluation in each side of the eye were also observed separately. Table 8.1 depicts the average variation *avg*, the standard deviation *std* and the percentage of images that presents the variation. The results were grouped by eye and side: right eye temporal (RET), right eye nasal (REN), left eye temporal (LET) and left eye nasal (LEN).

The results show that a certain variability occurs in the optometrist's evaluations and, therefore, this variability is expected to appear in the outputs of the system when comparing a pair of *reference* and *altered* subsets. However, it must be noted that, as the evaluations have a granularity level of 0.5, the average variation is low. Moreover, the majority of images do not show a variation in their automated evaluation.

Table 8.1: Variation of the experts grading in the same patient during different checkups.

Type	Checkups(1, 2)			Checkups(3, 4)		
	Avg	Std	% img affected	Avg	Std	% img affected
RET	0.129	0.280	20.00	0.157	0.265	28.57
REN	0.257	0.306	45.71	0.229	0.329	37.14
LET	0.171	0.296	28.57	0.171	0.296	28.57
LEN	0.243	0.351	40.00	0.171	0.241	34.29

8.2.2 Effect on the segmentation of the conjunctiva

The parameters for the automatic segmentation algorithms are set as in the previous experiments (Table 5.1).

The results of the application of the segmentation to both pairs of *reference* and *altered* sets is depicted in Table 8.2. Additionally to the sensitivity, specificity, accuracy and precision, the rate of false positives and false negatives is included to improve the comparison. The values are similar in all the cases, specially when the alteration is the presence of blue dye in the tears. The contact lenses have a bigger influence, as the largest differences are both the sensitivity and the percentage of false negatives on the contact lenses set. An example of the results of the segmentation procedure in pairs of images with and without alterations can be observed in Fig. 8.4.

Table 8.2: Validation of the repeatability of the ROI extraction procedure.

Set	Blue dye						Contact lenses					
	Sens.	Spec.	Accu.	Prec.	% FN	% FP	Sens.	Spec.	Accu.	Prec.	% FN	% FP
S_{ref}	0.875	0.835	0.836	0.853	0.079	0.085	0.833	0.883	0.834	0.905	0.113	0.053
S_{alter}	0.851	0.833	0.819	0.839	0.091	0.090	0.805	0.905	0.820	0.921	0.138	0.042

8.2.3 Effect on the feature computation

Once it has been proved that the alterations have a small influence in the image segmentation, the next step is to observe the changes in the values of the image features. To that end, the coefficient of variation and the difference between *reference* and *altered* sets is computed for each experiment. The difference in a given set S is calculated as:

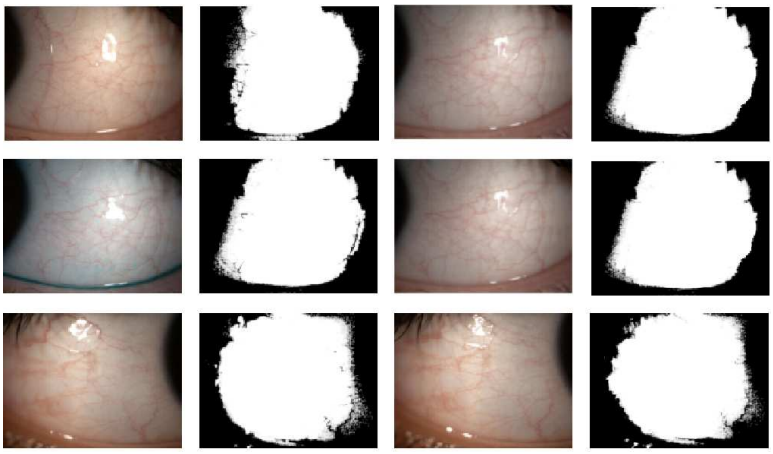


Figure 8.4: Pairs of images of the same side of the same eye that should produce a similar segmentation. The first pair shows the effect of the contact lenses, the second pair depicts the effect of the blue dye and the last pair shows two different optimal images.

$$diff_{subset} = \frac{S_{subset\_alter} - S_{subset\_ref}}{S_{subset\_ref}} \tag{8.2}$$

where S_{subset} is S_{blue} or S_{cont}. The results of the experiment are depicted in Table 8.3.

The results show that some of the features, such as I_1, remain stable through all the tests. However, most of them vary to different degrees depending on the image alteration. The features that present the smallest differences between subsets in the *blue* case, ordered from lowest to highest, are I_1, V_6 and V_1. The features that present the smallest differences between subsets in the *contacts* case, in the same order, are W_v, I_3, I_1, B_9, B_7, V_3, B_1, V_4, I_2 and B_3. Generally, the differences are lower in the *contacts* scenario. This happens due to the nature of the features, as most of them are colour-based and, therefore, will experiment higher changes when facing a variation of hue than adding contact lenses. In fact, A_v and P_v are the two features that present the highest differences for S_{cont}, and both of them are vessel-based. Also, the differences for these two features in S_{blue} are much lower.

The image features can be divided in four groups, depending on the variation that the feature can experience in each case, as depicted in Table 8.4. There are some features that remain within the same range when computed in the four subsets, which

Table 8.3: Coefficient of variation for each feature and differences between altered and reference sets.

Feature	Blue dye			Contact lenses		
	CV		diff(S_{blue})	CV		diff(S_{cont})
	S_{ref}	S_{alter}		S_{ref}	S_{alter}	
B_1	0.060	0.112	0.868	0.108	0.118	0.093
B_2	0.024	0.274	10.297	0.102	0.034	0.662
B_3	0.319	1.144	2.583	0.351	0.315	0.105
B_4	0.054	0.105	0.965	0.103	0.115	0.121
B_5	0.026	0.243	8.340	0.024	0.010	0.610
B_6	0.266	1.221	3.588	0.246	0.306	0.243
B_7	0.069	0.125	0.807	0.115	0.124	0.076
B_8	0.177	0.212	0.200	0.235	0.266	0.129
B_9	0.062	0.111	0.788	0.107	0.115	0.072
V_1	0.521	0.577	0.107	0.433	0.587	0.355
V_2	0.170	0.442	1.606	0.201	0.247	0.228
V_3	0.210	0.539	1.570	0.246	0.265	0.077
V_4	0.061	0.098	0.605	0.111	0.122	0.096
V_5	0.184	0.825	3.478	1.348	0.506	0.625
V_6	1.261	1.186	0.060	1.264	1.063	0.159
V_7	0.163	0.435	1.666	0.189	0.252	0.335
I_1	0.305	0.300	0.017	0.238	0.226	0.054
I_2	0.253	1.078	3.259	0.260	0.286	0.098
I_3	0.275	1.062	2.863	0.300	0.292	0.028
I_4	0.026	0.239	8.201	0.024	0.009	0.608
I_5	0.265	1.190	3.497	0.246	0.306	0.242
C_v	0.490	0.653	0.333	0.476	0.564	0.185
A_v	0.348	0.488	0.401	0.250	0.439	0.754
P_v	0.348	0.488	0.401	0.250	0.439	0.754
W_v	0.082	0.368	3.475	0.081	0.080	0.013

implies that their range of values is unlikely to be affected by image conditions. Examples of this are V_4, B_1, B_4, B_7, B_9, C_v, V_1 and V_6, hinting that features that are calculated using the hue in the background are less affected than the ones that take into account the vessels, an effect that is specially noticeable in S_{blue}. Therefore, these features are preferred, as their values are less affected by image alterations.

Table 8.4: Features grouped by coefficient of variation.

% of variation	Blue dye		Contact lenses	
	S_{ref}	S_{alter}	S_{ref}	S_{alter}
$\leq 20\%$	$V_2, I_4, V_4, V_5, V_7, B_1$	V_4, B_1, B_4, B_7, B_9	I_4, V_4, V_7, B_1, B_2	I_4, V_4, B_1, B_2
	$B_2, B_4, B_5, B_7, B_8, B_9$		B_4, B_5, B_7, B_9	B_4, B_5, B_7, B_9
$20\% - 30\%$	I_2, V_3, I_3, I_5, B_6	I_4, B_2, B_5, B_8	A_v, I_1, V_2, I_2, V_3	I_1, V_2, I_2, V_3
			P_v, I_5, B_6, B_8	I_3, V_7, B_6, B_8
$30\% - 40\%$	A_v, I_1, P_v, B_3		I_3, B_3	I_5, B_3
$\geq 40\%$	C_v, V_1, V_6	C_v, A_v, V_1, I_1, V_2	C_v, V_1, V_5, V_6	$C_v, A_v, V_1, P_v, V_5, V_6$
		I_2, V_3, I_3, P_v, V_5		
		V_6, I_5, V_7, B_3, B_6		

8.2.4 Effect on the training of the system

The results of the application of the feature selection techniques in IMG_2 dataset are depicted in Table 8.5. It can be observed how several of the chosen features vary their range depending on the experiment, as it was mentioned when describing the results of Table 8.4. Thus, noticeable differences are expected between the *reference* and *altered* subsets. Moreover, these differences are expected to be larger in the *blue* dataset, as most selected features are colour-based, with only one vessel-related feature, W_v, selected in the SMOReg approach.

Table 8.5: Features that appear in at least 7 out of 10 folds.

Method	#	selected features
CFS	12	V_1, V_2, I_2, V_3, I_4, I_5, V_7, B_2, B_5, B_6, B_7, B_9
Relief	8	I_1, I_3, I_4, V_6, V_7, B_2, B_3, B_5
M5	7	V_1, V_3, I_3, I_4, V_5, V_7, B_9
SMOReg	13	V_1, V_2, V_3, I_3, I_4, V_5, V_6, V_7, B_1, B_2, B_5, B_9, W_v

The MSE results for the machine learning techniques in each case are detailed in Table 8.6. The parameters for the regression methods are the same as the previous experiments, and can be seen in Table 7.1. The RF approach has low differences in both *blue* and *contacts* sets. The other two methods are more affected by the blue dye test, resulting in far poorer values in S_{blue} test than for S_{cont}. This is specially noticeable in PLS approach with all the features.

Table 8.6: MSE for each combination of features set and regression technique.

	S_{blue}			S_{cont}			IMG_2		
	MLP	RF	PLS	MLP	RF	PLS	MLP	RF	PLS
All	0.355	0.136	0.067	0.326	0.031	0.028	0.283	0.131	0.119
CFS	0.500	0.097	0.071	0.376	0.061	0.040	0.404	0.117	0.118
Relief	0.142	0.106	0.068	0.081	0.083	0.072	0.187	0.132	0.126
M5	0.366	0.084	0.076	0.248	0.046	0.038	0.310	0.118	0.118
SMOReg	0.173	0.095	0.069	0.109	0.045	0.037	0.234	0.120	0.118

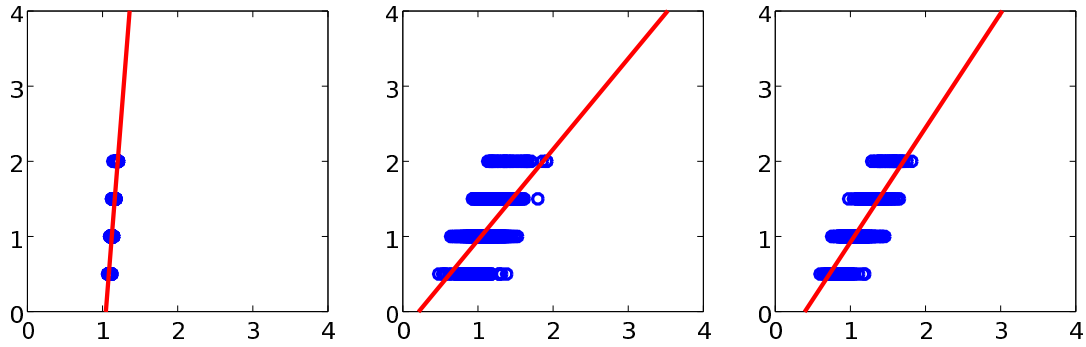


Figure 8.5: Scatter plots for each approach with their best feature subset and IMG_2 set. The x-axis and y-axis represent the predicted and real values respectively. Left to right: MLP with Relief, PLS with CFS and RF with SMOReg.

In order to have an idea of the general performance of the approaches, the MSE values for IMG_2 are also depicted in Table 8.6. The approaches that obtain the lowest MSE are the PLS with the CFS subset and the RF with the SMOReg subset. Both these approaches provide low differentiation in both tests and, thus, the regression techniques provide an evaluation close to the optometrist's.

In order to ensure the reliability of the results, other goodness metrics were also computed for each set (S_{cont}, S_{blue}, IMG_2): mean absolute error (MAE) and coefficient of determination (R^2). The best approaches according to the MSE obtained also the best results with MAE and R^2, with small changes in the order. For IMG_2 set, PLS with CFS obtained a $MAE = 0.272$ and $R^2 = 0.328$, while RF with SMOReg obtained a $MAE = 0.276$ and $R^2 = 0.590$. Finally, Figure 8.5 depicts the results for each regression technique with its best feature set (best values on the Table 8.6, IMG_2 set).

8.2.5 Effect on the final outputs of the system

Additionally, a statistical test was conducted in order to analyse if the differences between the automatic outputs and the optometrist's evaluations are significant. The experiment was conducted with the best approach, the RF with CFS. First, a normality test was performed in both samples. The null hypothesis was strongly rejected. As a consequence, a Wilcoxon signed rank test for the differences was conducted, as it does not require for the data to follow a normal distribution. The null hypothesis is that the differences come from a distribution with a zero median, and it was accepted with $\alpha = 0.05$ and a p-value of 0.115.

As the objective of the automatic methodology is to mimic the experts' behaviour, the variability of the specialist's evaluation between the checkups was compared to the one in the outputs of the automatic systems. As it was mentioned previously, the dataset IMG is obtained from a clinical study regarding contact lenses comfort. In this study, the differences among checkups were analysed, and reported as low. Therefore, it must be confirmed that the outputs from the automatic systems are within the same range. To that end, the outputs of the regression methods with the full feature set were obtained for all the images on the IMG dataset. The regression techniques' outputs were computed as described in Section 7.2, with 100 iterations of 10-fold cross-validation and taking into account only the validation set outputs in each fold. Then, the average differences between consecutive checkups were computed. This implies that, for a given combination of patient, eye and side, three differences in evaluations were computed: between checkups 1 and 2, between checkups 2 and 3, and between checkups 3 and 4. Moreover, these values were computed for the manual evaluations, as a means to establish the comparison. The results for the average difference and standard deviation are depicted in Table 8.7. Note that the coefficient of variation is not needed, as the compared values are within the same range. In fact, as the means are close to zero in some of the cases, and the observations may have different sign, the CV can be misleading.

In view of the data, the systems' outputs and the manual evaluations have a similar behaviour. The standard deviation is lower for the automatic approaches, and some

Table 8.7: Differences on the evaluation of the same case through different checkups.

System	MLP	PLS	RF	Manual
Avg. diff(C_1,C_2)	0.028	0.007	0.015	0.077
Avg. diff(C_2,C_3)	-0.079	-0.019	-0.044	-0.105
Avg. diff(C_3,C_4)	0.030	0.020	0.042	0.037
Std. diff(C_1,C_2)	0.355	0.233	0.257	0.467
Std. diff(C_2,C_3)	0.344	0.233	0.236	0.558
Std. diff(C_3,C_4)	0.365	0.237	0.246	0.528

of the systems also present a lower mean value in a given pair of checkups. In order to establish a deeper comparison, the cases that the optometrist labelled with the same value in a pair of checkups were analysed. The number of images that fulfil the condition in the checkups 1-2, 2-3 and 3-4 was 96, 92 and 92 respectively. The variation of the automatic outputs is shown in Table 8.8. The average differences of the systems are close to zero for all the cases. The three systems present a similar behaviour, and the best one varies depending on the pair of checkups. For the pair (C_1,C_2), the best system is the PLS approach, while the RF offers the best results for the other two cases. The MLP presents the highest standard deviation, and is the worst method overall.

Table 8.8: Magnitude of the variation in the automatic systems for those cases where the manual evaluation does not vary.

System	MLP	PLS	RF
Avg. diff(C_1,C_2)	-0.035	-0.014	-0.020
Avg. diff(C_2,C_3)	0.028	0.056	0.048
Avg. diff(C_3,C_4)	-0.027	-0.002	-0.001
Std. diff(C_1,C_2)	0.234	0.163	0.204
Std. diff(C_2,C_3)	0.231	0.215	0.201
Std. diff(C_3,C_4)	0.238	0.209	0.189

8.3 Conclusions

In this chapter, the influence that different image alterations have in each step of the automatic methodology is analysed. The inputs of the system have a high variability,

and there are some scenarios where, although the images are visually different, the optometrists' evaluations do not vary. Therefore, in order to ensure that the automatic system is able to mimic the experts' behaviour, the reaction of manual and automatic evaluations under two image alterations was measured and compared.

The segmentation of the region of interest is more affected by the presence of contact lenses, as the edges and reflections can hinder the performance of the segmentation algorithms. However, the segmented region only suffered from small variations in the tests. The feature computation is more affected by the remains of blue dye used to detect staining, as most of the image features are colour-based and, therefore, present larger variations when the hue of the conjunctiva varies. The features that took into account only the background of the conjunctiva were more stable than those that included the vessels. Finally, regarding the regression techniques that transform the image features to grading scale values, the approach that obtained the best results through the different tests was the RF. Moreover, when comparing the magnitude of the changes of the systems when the experts' evaluations do not vary, the three methods obtained accurate results.

In view of the obtained results, it can be concluded that the automatic methodology behaves like the human expert, and that noticeable alterations in the inputs that do not cause changes in the optometrists' evaluations do not cause changes in the automatic outputs either.

Chapter 9

Class imbalance problems

The highest and lowest values of the grading scales (Fig. 9.3) are very unlikely to happen. The former because patients tend to seek treatment before reaching that level of hyperaemia, and the latter because even healthy individuals tend to present traces of redness. Thus, in practice, specialists rate between the second level and the next-to-last one instead of the real boundaries established by the scale. Therefore, most of the images and videos of the data sets are evaluated as belonging to intermediate hyperaemia levels. For example, experts tagged more than 30% of the images from VID dataset about 2.0 in the Efron scale. The percentage rises above 40% for the intermediate value of the BHVI scale, 2.5. This causes a class imbalance problem, as there are few samples of extreme levels.

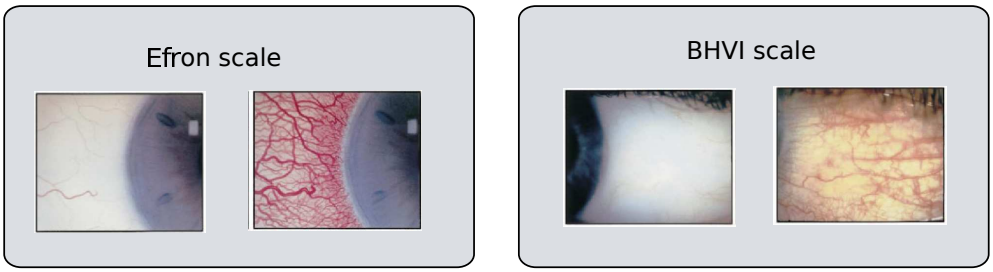


Figure 9.1: Lowest and highest prototypes of the grading scales. Left: Efron scale. Right: BHVI scale.

Moreover, healthy individuals can present a level of hyperaemia close to 2.0 [109], which implies that a healthy eye may not be as white as the grading scales show. Scales

are also imbalanced, as there are virtually no individuals whose eyes are in the lowest level. Most eyes have some degree of hyperaemia [110, 111], which also strengthens the idea that a real low hyperaemia value corresponds to an intermediate value in grading scales. For example, the image in Fig. 9.2 was tagged as 1.5 in the Efron scale. In these works, it is also concluded that grading scales are not linearly incremental, but more similar to a quadratic distribution [110].

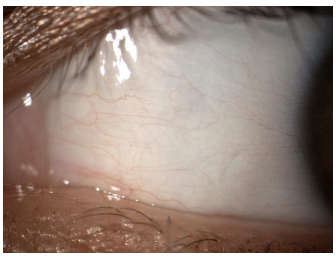


Figure 9.2: Example of one of the images tagged with a low hyperaemia value.

Finally, there are differences depending on the scale. For example, the distribution of values for the Efron and BHVI scales is not the same. As the Efron scale consists of drawings, the separation of the intervals is more even than the differences in BHVI.

In this work, the final output of the automatic methodology is computed by means of machine learning techniques. In order to succeed, these techniques benefit from an even distribution in the number of instances through all the classes. On the contrary, when they are applied to an imbalanced dataset, several problems may arise. On one hand, the trained system will not be able to recognise the instances of the classes that have a low number of elements. On the other hand, when computing quality metrics to assess the behaviour of the system, the classes with few instances are easily ignored, as their contribution is minimal.

The automatic methodology proposed in this work has been developed and validated with imbalanced datasets. Therefore, the obtained results are limited by that suboptimal data distribution. Thus, a reasonable question arises: how will the methodology react if the dataset was closer to an ideal, more balanced distribution? The objective of this chapter is to answer that question by analysing an imbalanced image set and applying the necessary techniques to solve or at least minimise the problem.

To that end, the selected dataset was VID_1. This dataset is imbalanced, as the distribution of values is uneven, with most of the images falling in the middle class of the grading scales, as depicted in Fig. 9.3. Moreover, it is a small dataset, so it can benefit from the artificial generation of extra samples. Finally, it is evaluated in both the Efron and the BHVI grading scales, allowing the comparison of the effect that the proposed techniques have in several scales.

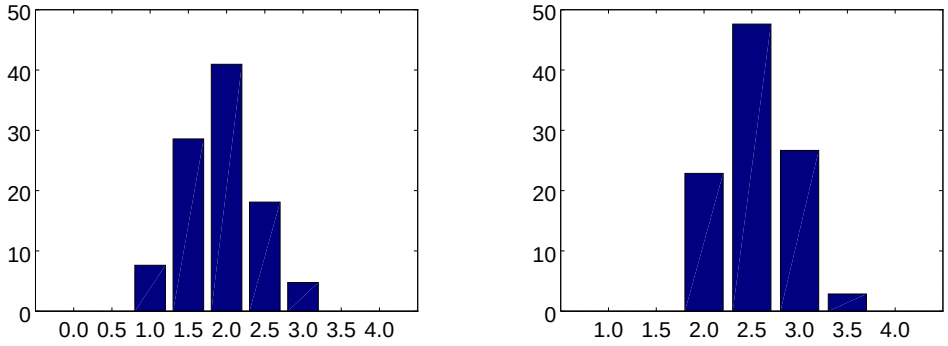


Figure 9.3: Distribution of values in VID_1 dataset. y-axis represents the percentage of samples that are labelled as belonging to each class. Left: Efron scale. Right: BHVI scale.

The tests in this chapter were tackled using the complete methodology. That is, the inputs of the system were the videos from the VID_1 dataset, that went through the frame selection step. Then, the conjunctiva segmentation step applied the approach M_{ET} from Chapter 5, as it provides good results with VID dataset. However, the images that obtained a poor segmentation were removed in order to ensure that the segmentation would not affect the final outputs. Therefore, only 105 of the 114 images of the VID_1 dataset were used. This test was performed taking only the 25 global features into account.

9.1 Data balancing methods

In order to tackle the class imbalance problem, it is necessary to modify the distribution of values of the data set. There are several approaches in the literature [112, 113], and in this chapter, three of the most widely used were applied:

Undersampling removes values from the largest class. The values to be removed can be selected randomly or taking their information into account. In this chapter, the values furthest from the class prototype are removed.

Oversampling generates additional values for the smallest class. There are two approaches, either replicating an existing value or artificially generating a new one. In this chapter, the values closest to the class prototype were replicated.

Synthetic Minority Over-sampling TEchnique (SMOTE) combines both undersampling of the largest class and oversampling of the smallest. The new values are artificially generated by computing the k nearest neighbours for each element of the class and selecting one of them. Two experiments were performed with different oversampling percentages.

As a first step to the application of any of the methods, the experts' evaluations must be divided in classes. There are several possible partitions of the data. The minimal precision that optometrists used for the gradings was 0.1, therefore this is the minimal difference between classes that can be considered. However, this partition is too thin, and the data set do not have enough elements to represent all the prototypes. A value above 1.0 is too inclusive, as the data will be divided in fewer classes than there are available scale prototypes. Studies in the literature [15, 14] have depicted how the human experts, even if they have a wider range of values available, tend to grade their patients with integer or half integer values. Therefore, these are the steps that were used in this study.

9.2 Class splitting

This section will detail the splitting in classes with $step = 0.5$ for each method and how each approach affects to some of the regression methods proposed in Chapter 7.

The most common class in VID_1 has only 43 and 50 samples in the Efron and BHVI scales, respectively. Thus, the SMOTE approach was applied without undersampling in the larger class, following the criteria of obtaining a similar number of images in all

the classes. Given that the distribution of the dataset is close to a normal function, it was decided to use a higher percentage on the most extreme classes, and a lower one for the classes closer to the central one (the most common). It must be noted that the objective of this experiment is to prove that the SMOTE method can reduce the class imbalance problem, rather than obtain the optimal configuration. Therefore, the two following configurations were tested:

- SMOTE 1:
  - Efron scale: classes 1 and 3, 300%; classes 1.5 and 2.5, 100%
  - BHVI scale: classes 2 and 3, 100%; class 3.5, 300%
- SMOTE 2:
  - Efron scale: classes 1 and 3, 500%; classes 1.5 and 2.5, 200%
  - BHVI scale: classes 2 and 3, 200%; class 3.5, 500%

Tables 9.1 and 9.2 depict the number of elements in each class for each method when the evaluations are splitted in classes using half-integer values as prototypes.

Table 9.1: Number of samples in each class using integer and half integer as prototypes (Efron scale).

Class	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	Total
# original	0	0	8	30	43	19	5	0	0	105
oversampling	0	0	43	43	43	43	43	0	0	215
undersampling	0	0	5	5	5	5	5	0	0	25
SMOTE 1	0	0	32	60	43	38	20	0	0	193
SMOTE 2	0	0	48	90	43	57	30	0	0	268

Table 9.2: Number of samples in each class using integer and half integer as prototypes (BHVI scale).

Class	1.0	1.5	2.0	2.5	3.0	3.5	4.0	Total
# original	0	0	24	50	28	3	0	105
oversampling	0	0	50	50	50	50	0	200
undersampling	0	0	3	3	3	3	0	12
SMOTE 1	0	0	48	50	56	12	0	166
SMOTE 2	0	0	72	50	84	18	0	224

9.3 Results

The experiments were performed with three of the regression methods analysed in Chapter 7 and four of the feature selection techniques proposed in Chapter 6. As the objective of this experiment is to study if balancing techniques can improve the results, it is irrelevant if the error obtained with the machine learning techniques is optimal. Therefore, three state-of-art techniques were picked from the available ones without taking into account their previous performances.

For comparison purposes, Table 9.3 depicts the obtained values with the systems trained and tested with the VID_1 image set. It must be noted that there is a discrepancy with the results obtained in Chapter 7 for the same regression systems. This happens because the results in this chapter have been computed applying the complete methodology, while results in Chapter 7 were computed starting from a manual segmentation in order to prevent the bias from previous steps masking the regression results. The discrepancy is about 20% in several cases. However, there are combinations where it is more noticeable, such as the MLP in the Efron scale and M5 feature set, that changes its MSE from 0.234 to 0.861.

Table 9.3: Comparison of MSE values for the MLP, RBFN and RF regression methods for features appearing in 7 out of 10 folds.

Method	MLP		RBFN		RF	
	Efron	BHVI	Efron	BHVI	Efron	BHVI
All features	0.218	0.137	0.219	0.140	0.112	0.074
Relief	0.218	0.137	0.221	0.140	0.133	0.077
CFS	0.108	0.137	0.219	0.140	0.125	0.076
M5	0.861	0.061	0.215	0.141	0.132	0.084
SMOReg	0.114	0.075	0.220	0.137	0.120	0.085
Combination	0.218	0.137	0.219	0.140	0.126	0.073

As the dataset consists of only 105 images, alternatives that apply oversampling will be preferred as undersampling removes too much information. The MSE values for both approaches are depicted in Tables 9.4 and 9.5.

Regarding the oversampling approach, nor the MLP neither the RBFN approaches are able to improve the results obtained with unaltered data. However, the RF approach

Table 9.4: MSE values for oversampling.

Method	MLP		RBFN		RF	
	Efron	BHVI	Efron	BHVI	Efron	BHVI
All features	0.161	0.172	0.352	0.151	0.063	0.054
Relief	0.124	0.211	0.370	0.155	0.084	0.075
CFS	0.243	0.246	0.327	0.154	0.072	0.069
M5	1.366	0.434	0.359	0.147	0.087	0.073
SMOReg	0.268	0.083	0.300	0.156	0.081	0.073
Combination	0.431	0.214	0.286	0.158	0.071	0.055

Table 9.5: MSE values for undersampling.

Method	MLP		RBFN		RF	
	Efron	BHVI	Efron	BHVI	Efron	BHVI
All features	0.453	0.157	0.498	0.168	0.419	0.063
Relief	0.482	0.159	0.472	0.158	0.527	0.093
CFS	0.496	0.073	0.594	0.168	0.485	0.084
M5	1.831	0.053	0.660	0.166	0.509	0.097
SMOReg	0.088	0.082	0.548	0.155	0.545	0.090
Combination	0.475	0.153	0.397	0.165	0.428	0.081

is able to exceed the values with the initial data for both scales. The MSE is reduced from 0.11 to 0.06 and from 0.07 to 0.06 in the Efron and BHVI scales, respectively. Undersampling, on the other hand, worsens the results in all the cases due to an insufficient number of instances.

Finally, the values obtained for the first configuration of the SMOTE approach are shown in Table 9.6. In comparison with the results obtained with unaltered data, the MLP improves some feature sets in both scales, such as the M5 wrapper in the Efron scale, although it worsens others considerably, such as the SMOReg wrapper also in the Efron scale. The RF approach was found to improve greatly in both scales. In contrast, the RBFN obtains worse results.

The results for the second configuration of the SMOTE approach are depicted in Table 9.7. When comparing the two SMOTE configurations, the results for the RBFN and RF approaches are similar, with the exception of RBFN with the SMOReg set and the Efron scale, that achieves a much lower error values. The RF approach obtains a

Table 9.6: MSE values for the first configuration of the SMOTE approach.

Method	MLP		RBFN		RF	
	Efron	BHVI	Efron	BHVI	Efron	BHVI
All features	0.334	0.009	0.335	0.183	0.032	0.025
Relief	0.011	0.010	0.332	0.186	0.027	0.025
CFS	0.010	0.182	0.332	0.184	0.030	0.025
M5	0.012	0.011	0.337	0.185	0.030	0.025
SMOReg	0.330	0.183	0.332	0.188	0.029	0.024
Combination	0.329	0.010	0.331	0.184	0.029	0.025

slightly lower error in the second configuration in most cases. Finally, the MLP is again slightly inconsistent, improving some of the approaches but worsening others.

By comparing SMOTE with oversampling or undersampling approaches, it can be observed how the SMOTE configurations offer achieve lower MSE values for both MLP and RF. However, the best values for the RBFN are generally achieved with oversampling, as SMOTE affects this method badly.

Table 9.7: MSE values for the first configuration of the SMOTE approach.

Method	MLP		RBFN		RF	
	Efron	BHVI	Efron	BHVI	Efron	BHVI
All features	0.348	0.212	0.348	0.212	0.022	0.024
Relief	0.349	0.008	0.347	0.212	0.021	0.022
CFS	0.010	0.007	0.350	0.213	0.021	0.023
M5	0.006	0.010	0.351	0.213	0.020	0.021
SMOReg	0.348	0.212	0.035	0.212	0.022	0.024
Combination	0.010	0.211	0.352	0.213	0.022	0.022

9.4 Conclusions

This chapter is focused on a problem that appears commonly in methodologies that receive real-world information as input: the absence of a large database and the uneven distribution of samples. It is not uncommon that, on novel research scenarios, the databases are scarce, which forces researchers to work with reduced sets of samples that are not able to provide enough generalisation capability. Moreover, in environments such as bulbar hyperaemia grading, the images available did not cover the full range of

values displayed in the grading scales. This poses a problem, as the machine learning techniques cannot learn something without seeing it.

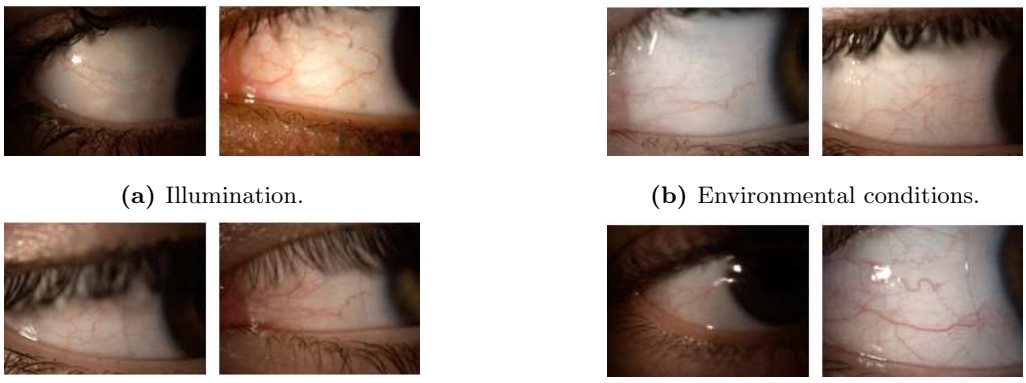
To solve these issues, data balancing approaches were analysed. These methods are used to increment the number of observations while ensuring that the number of samples in each class is similar. In view of the results, the SMOTE method can effectively reduce this issue, although the absence of values for extreme classes will remain unsolved. The best results for the Efron scale are achieved with the SMOTE approach with the oversampling percentages 500% in classes 1 and 3 and 200% in classes 1.5 and 2.5. The best feature selection technique for this scale was the M5, followed by the CFS and the combination of all the features. Moreover, the oversampling approach also obtains good MSE values in both the MLP and RF systems with several feature selection sets. Finally, regarding the BHVI scale, the MSE values tend to be lower. In this case, the SMOTE approach with the MLP obtain the best values.

Therefore, it has been proved that the balancing techniques can improve the results obtained with the automatic approach for hyperaemia grading in the bulbar conjunctiva. This has a strong repercussion in the capabilities of the methodology, as it can be applied even in those environments where the data capture is difficult, or in the first stages of an experiment, when there are few samples available. However, in the end these techniques can only be as effective as the dataset allows them to be, as there are minimal requirements of information that only an improvement in the quantity or quality of the images can overcome.

Chapter 10

Precise segmentation

Obtaining an accurate segmentation of the conjunctiva has proven to be a convoluted task. The process is hindered by several image issues, such as the variations in illumination, the characteristics of each device used to take the images or videos, the environmental conditions, the position of the eye in the image (close to the camera, centred or not), or the percentage of eye openness. Figure 10.1 depicts examples of the different situations aforementioned. Moreover, the segmentation can be a time-consuming stage in the automatic methodology for hyperaemia grading.



(a) Illumination.

(b) Environmental conditions.

(c) Quality of focus.

(d) Proximity of camera to conjunctiva.

Figure 10.1: Variability in the image set.

There is a wide range of possibilities in real-world environments, all of them equally valid for bulbar hyperaemia grading. Nevertheless, because of the non-standardised

capture process, there are only certain areas of the conjunctiva that appear in all the images. Therefore, in order to ensure the applicability of the methodology, it is necessary to study if a reduced area can provide enough information to compute the hyperaemia evaluation.

The optometrists say that they look at the whole conjunctiva when performing the grading. However, the variability of the images shows the opposite, as there are several images that offer a good depiction of the conjunctiva but some areas (mainly near the eyelids) are missing. This hints at the possibility that some regions are more relevant than others. Moreover, the results obtained in the study that assess the influence of the local features support this idea, as the features' relevance varies depending on the part of the eye where they are obtained. Thus, the specialists may use the whole area, but they do not use it evenly.

This knowledge is difficult to model even for the optometrists themselves. Therefore, the only way to reach a conclusion on the matter is to perform an exhaustive and objective experiment regarding different areas of computation. To that end, the computation of the features is restricted to the central part of the image, an approach supported by Rodriguez *et al.* [18], where the region of interest consists in a small rectangle that is manually segmented in the central area of the image. Additionally, in the study by Yoneda *et al.* [17] a rectangular region is manually defined in the image, which is used to analyse the influence of the number of vessels in bulbar hyperaemia.

In this chapter, the bulbar conjunctiva is divided into smaller regions of interest. Then, the contribution of each area in the computation of hyperaemia is studied. To that end, the 25 features defined in Chapter 6 are computed in each of the sub-areas, and a feature selection approach is applied in order to establish which ones are the most relevant. Then, by means of regression methods, the sets of features are transformed to the grading scale, and the results are analysed and compared to the values obtained within the whole conjunctiva.

10.1 Defining a suitable region of interest

The dataset IMG'_1 is used for this study. This dataset was selected because the size of the conjunctiva was generally less variable. Moreover, the eye tends to be closer to the camera than in VID dataset, which allows the segmentation of larger sections. This dataset consists of 76 images of the bulbar conjunctiva obtained from the IMG set, that have been evaluated by two specialists in the Efron scale. The images where the experts' gradings vary more than 0.5 were discarded.

A manual segmentation was performed for each of the images in order to limit the posterior stages of the computation to the conjunctiva area. By using this manual segmentation as a starting point, a rectangle of 512×512 px was defined in the centre of each mask, as shown in Fig. 10.2. The particular conditions of the dataset guarantee that the central area of the image, where the rectangle is selected, is mostly composed of conjunctival pixels. This happens because the pictures were taken in a similar fashion, and the optometrist had the objective of capturing most of the conjunctiva. Additionally, the remains of eyelids and eyelashes that get included in the rectangle can be removed by means of the manual mask, ensuring that they will not add bias in the results. The same principle applies for the iris area.

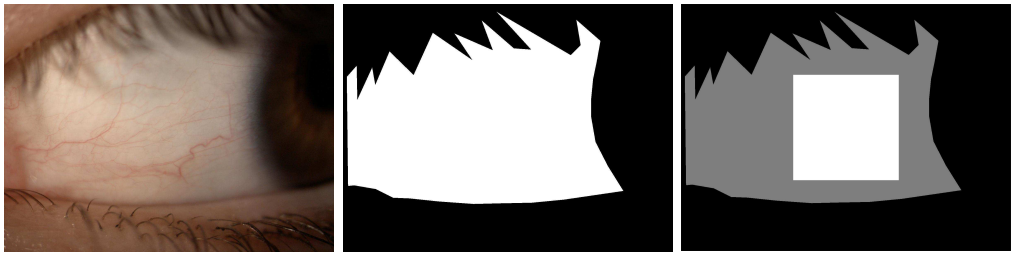


Figure 10.2: Conjunctiva image, manual segmentation of the region of interest and central square of 512×512 px.

The size of the area was decided as the larger central area that was present in most of the images, as the changes in the position of the conjunctiva within the image and the variability of eyelids and eyelashes prevent larger areas to be taken into account. Previous works [17] support the opinion that even smaller zones are relevant for the grading. Additionally, the selected area is required to be centred, as the objective is

to study the same region in all the images, and the central part of the conjunctiva is the only one that appears in all the images, once again due to the variability of the dataset. Nevertheless, six of the analysed images did not provide an area large enough, because the eye was slightly closed. Therefore, these images were discarded.

The central 512×512 square was divided into cells. Among many division possibilities, in this test the 1×2, 2×1 and 2×2 grids were considered (Fig. 10.3), as a region smaller than 256×256 px was deemed as insufficient to provide the necessary information for the evaluation.

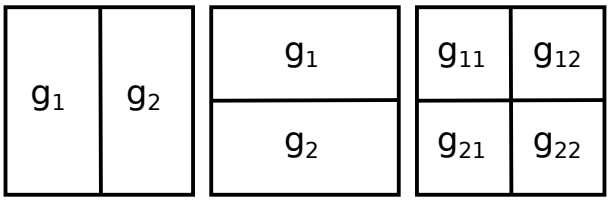


Figure 10.3: The three grid configurations used in the experiment.

Since some tests show differences between the image features in the iris area and in the corner of the eye/caruncle area, some of the images were flipped around a vertical axis. This way, all the images had the iris on the same side, allowing the comparison of the exact same area of the image in all the cases.

10.2 Results

The image features were computed for the 70 images in the IMG'_1 dataset where a feature vector for each configuration (1×2, 2×1 and 2×2) was obtained. The next step was to apply feature selection techniques to each case. Three feature selection methods methods were chosen: CFS, Relief and SMOReg. As in the previous experiments, 10-fold cross-validation was used, and the features selected were the ones that appeared in at least 7 out of 10 folds. For the Relief method, the features in the first ten positions of the rank were taken into account. Table 10.1 depicts the feature selection results.

In the table, the superscript s is used to label the features computed in the central square, while c is used for the features computed in the whole conjunctiva. The superscripts used to define each grid and position are the ones defined in Fig. 10.3.

Table 10.1: Features chosen with each grid and feature selection method.

Grid	CFS
1×2	A_v^s, P_v^s, I_5^s, B_8^s, P_v^c, B_8^c, B_8^1, A_v^2
2×1	A_v^s, P_v^s, I_5^s, P_v^c, B_8^c, B_8^1, B_8^2
2×2	A_v^s, P_v^s, I_5^s, P_v^c, B_8^c, B_8^{11}, B_8^{21}, A_v^{22}
Grid	Relief
1×2	I_5^s, B_6^s, V_7^s, V_2^s, I_2^s, V_3^s, V_2^c
2×1	I_5^s, B_6^s, V_7^s, V_2^s, I_2^s, V_2^c, V_7^2, V_3^s
2×2	I_5^s, B_6^s, V_7^s, V_2^s, I_2^s, V_3^s, V_2^c, V_4^s
Grid	SMOReg
1×2	P_v^s, I_5^s
2×1	I_5^s, B_2^2
2×2	I_5^s

As expected, the methods favour larger areas, both the central square and the whole conjunctiva, as they provide more information than the smaller cells. However, there are exceptions to that tendency, and each method chooses at least one feature computed in a part of the grid for the final subset. That is the case of feature B_8 in CFS, feature V_7 in Relief, and feature B_2 in SMOReg.

The logical question that arises at this point is whether or not it is possible to perform an accurate evaluation of bulbar hyperaemia by using only the features computed in smaller regions of the eye. Therefore, an additional experiment was performed by applying feature selection including only the features computed in the cells. The procedure is the same than in the previous experiment, and the results are depicted in Table 10.2.

By comparing the features chosen in both tables, some similarities are noticeable, such as the occurrence of features P_v and B_8, that remain being favoured by the techniques. However, it is interesting that features such as I_5, which was considered relevant in most of the cases in the previous experiment, is no longer selected here. This implies that the feature is probably more relevant when it takes place in a large area, but not in smaller ones. The opposite situation occurs with feature V_5, that appears more often.

Analysing the relevance of each region, in the 1×2 and 2×1 configurations both CFS and Relief seem to chose features indistinctly in both areas. However, differences

Table 10.2: Features chosen with each grid and feature selection method (cells only).

Grid	CFS
1×2	P_v^1, B_8^1, A_v^2, P_v^2
2×1	P_v^1, B_8^1, V_2^2, P_v^2, V_7^2
2×2	P_v^{12}, A_v^{21}, P_v^{21}, V_7^{21}, B_8^{21}, A_v^{22}, P_v^{22}
Grid	Relief
1×2	I_5^1, V_2^1, I_5^2, B_6^1, V_3^1, V_2^2, I_2^1, B_6^2, V_7^1, I_2^2, V_3^2, V_7^2
2×1	V_7^2, I_5^2, B_6^2, V_3^2, I_2^2, V_2^1, I_3^2, V_3^1, V_7^1, B_3^2
2×2	I_5^{21}, V_2^{21}, B_6^{21}, V_3^{21}, I_2^{21}, I_5^{12}, V_7^{21}, B_6^{12}, I_3^{21}, I_5^{22}, I_2^{12}
Grid	SMOReg
1×2	V_5^1
2×1	V_7^2
2×2	V_5^{21}, I_5^{21}

appear in SMOReg approach, as it favours the inferior and left areas of the eye. In the configuration with the smallest cells, 2×2, SMOReg selects features only in the bottom left corner, while the other methods select features from all cells but g_{11}, which implies that the upper left corner is the least relevant.

Three of the regression techniques that obtained the best results were trained for each combination of grid and feature selection technique, and the cross-validation MSE was obtained. In order to facilitate the comparison, a row with the results of the features selected in the manually segmented conjunctiva for the same images was included. The results are depicted in Table 10.3. The MLP obtains the lowest MSE for all the configurations. Both configurations of 2 cells achieve the best value with all the features. For the 2×2 configuration, the best feature set is the SMOReg one, that consists in only feature I_5 computed in the whole conjunctiva.

Regarding the experiments that take into account only the features computed in the individual cells, the results are depicted in Table 10.4. The best value obtained for the 1×2 configuration is once again obtained by the MLP with all the features. Still, its MSE is interesting, as it improves the original minimal MSE obtained by the feature vector that included the features in the whole conjunctiva and in the 512×512 square. The MLP also obtains the best results for the 2×2 configuration with the CFS subset. Finally, the PLS approach obtains the lowest MSE in the 2×1 configuration with the

Table 10.3: MSE obtained for the features chosen with each grid and feature selection method (both cells and global features). The best value for each configuration is highlighted.

			MLP	
Grid	All	CFS	Relief	SMOReg
1×2	0.022	0.221	0.048	0.045
2×1	0.030	0.039	0.046	0.054
2×2	0.221	0.045	0.030	0.030
Conjunctiva	0.221	0.223	0.221	0.057
			PLS	
Grid	All	CFS	Relief	SMOReg
1×2	0.072	0.088	0.053	0.053
2×1	0.058	0.114	0.054	0.064
2×2	0.072	0.140	0.052	0.061
Conjunctiva	0.064	0.053	0.055	0.054
			RF	
Grid	All	CFS	Relief	SMOReg
1×2	0.083	0.080	0.096	0.079
2×1	0.090	0.081	0.093	0.108
2×2	0.086	0.080	0.092	0.102
Conjunctiva	0.083	0.102	0.097	0.109

Relief feature set. These two values do not improve the previous minimum MSE for the given grid, but they still achieve a value lower than 0.1. As it was mentioned, is not uncommon that the optometrists differ in more than 0.5 for the same image, that is, in a squared error higher than 0.25. Therefore, the results support that the system that takes into account only the reduced region of interest also behaves like a human expert. Moreover, these results imply that the automatic methodology can provide accurate results even when only part of the conjunctiva is available.

Table 10.4: MSE obtained for the features chosen with each grid and feature selection method (cells only). The best value for each configuration is highlighted.

			MLP	
Grid	All	CFS	Relief	SMOReg
1×2	0.021	0.100	0.221	0.102
2×1	0.221	0.221	0.221	0.353
2×2	0.221	0.055	0.221	0.058
			PLS	
Grid	All	CFS	Relief	SMOReg
1×2	0.071	0.117	0.086	0.239
2×1	0.069	0.294	0.068	0.070
2×2	0.095	0.096	0.078	0.071
			RF	
Grid	All	CFS	Relief	SMOReg
1×2	0.093	0.146	0.100	0.251
2×1	0.100	0.100	0.109	0.108
2×2	0.103	0.112	0.105	0.130

10.3 Conclusions

In this experiment, the fully automatic methodology for hyperaemia grading is used to identify the most relevant areas of interest in the bulbar conjunctiva.

There are two main objectives in this chapter. On one hand, to analyse if a small area of the conjunctiva can be representative enough for grading purposes, as the variability of the images frequently prevents the retrieval of the complete region. On the other hand, to identify in which areas a given feature is more relevant, and what regions have most of the optometrists' interest.

To that end, a central square of the image was selected. The central region is the most constant section among the images, as it is present despite of the position of the camera in relation to the eye or the quantity of eyelids and eyelashes that are depicted. The selected area is the largest one that was allowed without including noticeable spurious regions. This central square was then sub divided into cells in order to analyse the effect of smaller areas.

As the experiment seeks to determine which features are more relevant in each area, feature selection techniques are applied to the original sets. The results illustrate how some features are indicators of hyperaemia only if they take place in large areas, in opposition to others that only gain relevance in the smallest cells.

The next step is to establish the correspondence between each feature set and the values in the Efron grading scale. Three of the regression techniques that obtained good results through past experiments were applied. For the configurations using both global (conjunctiva and central rectangle) and local (cells) features, the best values were obtained by the MLP with all the features in the 1×2 grid. Still, this value can be improved by using only local features. Furthermore, there are several cases where cells-only approaches are able to obtain lower error results than some of the global and local combined cases.

Thus, it can be concluded that the hyperaemia in the bulbar conjunctiva can be measured as long as the central area of the image is available, as it is representative enough. Therefore, the most important feature of a segmentation approach is to represent clearly this central area without including spurious regions. This has potential repercussions in a reduction of the computational time invested and a lower chance of including non-relevant information. Moreover, the fact that only certain regions are essential implies that the automatic methodology can still be effective in images that show a smaller region of the conjunctiva, or low quality images where only certain areas can be used, such as when large shadows are present near the edges.

Finally, the experiment with only locally-computed features points at the lower region of the eye being more important than the upper one, and the iris side being more relevant than the opposite one.

Appendix A

Materials and methods

All the experiments were conducted on a Intel Core 2 Quad CPU (2.83 GHz) and 4 GB of RAM. The operative system was Debian version 3.2.57 kernel 3.2.0. This appendix includes additional clarification on the technologies that were used during the work.

A.1 OpenCV

OpenCV [114] (Open source Computer Vision library) is a computer vision library that has interfaces for several programming languages. OpenCV is cross-platform (Windows, Linux, MacOS, iOS, Android) and includes classic and state-of-the-art computer vision and machine learning algorithms. The library is widely used in companies and research groups. It was originally developed by Intel and currently has a BSD license, so it is free for both academic and commercial use.

The C++ programming language and the OpenCV library were used during the first three steps of the proposed methodology. Specifically, they were used to implement the illumination and blurriness for the frame selection, all the segmentation approaches for the conjunctiva extraction and the computation of the image features. The version of OpenCV used was 3.0.0, with gcc version 4.6.3.

A.2 Matlab

Matlab [115] (MATrix LABoratory) is an integrated development environment (IDE) that uses its own programming language. It is focused on mathematics, but it has a number of toolbox for different areas. In this work, the Neural Networks Toolbox and the Statistics and Machine Learning Toolbox have been used. The former offers implementations on most state-of-art artificial neural networks, while the latter includes regression techniques and statistic tools. Matlab is a cross-platform software, and versions for Windows, Linux and MacOS exist.

Matlab was developed by MathWorks, and is under a proprietary license.

Matlab was used during the fourth step of the methodology to implement the classifiers and regression techniques. Moreover, the manual segmentation of the conjunctiva that served as gold standard was also computed in Matlab. The version of the IDE was R2014a.

A.3 Weka

Weka [116] (Waikato Environment for Knowledge Analysis) is a software that comprises a large array of machine learning algorithms. It also includes visualisation and data analysis tools. Weka is portable, and can be executed in any system that runs Java.

Weka was developed in the University of Waikato (New Zeland), and it is licensed under GNU General Public License. Thus, the software can be run, studied, modified and shared at will.

Weka was used during the third step of the methodology to implement the feature selection techniques. Moreover, the comparison of classifiers and regression systems was also developed in Weka. Version 3.6.12 was used.

Appendix B

Colour spaces

A colour space is a means of specifying a colour in a concise and unique manner. There are several colour spaces that are used depending on the problem that is being solved or the environment circumstances. An image can usually be directly transformed from one colour space to another, but not always in a linear way. Moreover, some colour spaces can represent a wider range of colours than others. In this work, five of them are used: RGB, HSV, HSL, L*a*b* and TSL. A brief insight on each one is offered in this appendix.

B.1 RGB colour space

RGB colour space [117] is commonly used in computer-based applications. It consists of three channels, each one representing colours red, green and blue. Therefore, the possible combinations of channels are commonly represented as a cube, as depicted in Fig. B.1. These three channels have the same range, although different range of values can be used, depending on the representation. In this work, the range 0-255 was used. The diagonal of the cube represents the grayscale colours, from black to white.

As the input images of the system are read in RGB colourspace, in this work a transformation was applied from RGB to each of the other colourspaces used. As some of the transformations may vary, the remaining sections of this appendix will depict the formulation to transform RGB to these other colourspaces.

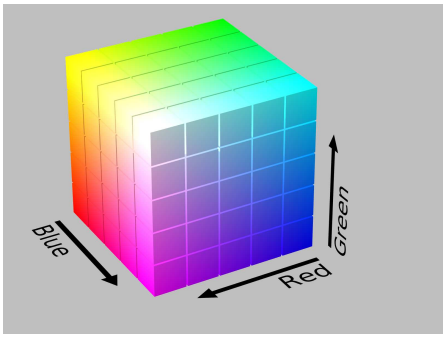


Figure B.1: Cubic representation of the RGB colourspace. Each axis represents one of the channels.

B.2 HSV and HSL colour spaces

HSL [118] stands for hue, saturation and lightness (or luminance). HSV [119] stands for hue, saturation and value.

The hue represents the same concept and values in both: an angle that ranges from 0 to 360 degrees. The six main colours are separated 60 degrees each, starting in red at 0, then yellow at 60, green, cyan, blue and magenta. The saturation ranges from 0 to 1, and sometimes is depicted as a percentage. It indicates how *dark* is the colour, with 1 being the pure colour and 0 representing black. However, it must be noted that the saturation channel, while representing the same concept, is not equivalent in both colour spaces.

HSL's lightness is also expressed in the range 0-1 and, intuitively, changes the illumination from low to high. That is, the colour will look more vivid with a low lightness value, and paler and closer to white with a high lightness value. HSV's value level varies the colour saturation, a 0 level represents black, and 1 represents the saturated colour.

HSV and HSL are commonly represented with cylindrical coordinates as depicted in Fig. B.2. They move from RGB's cubical representation in an attempt to be more intuitive and perceptually relevant.

HSV and HSL were developed for computer graphics applications, and they are widely used in image editing. However, they are not perceptually uniform. As they are

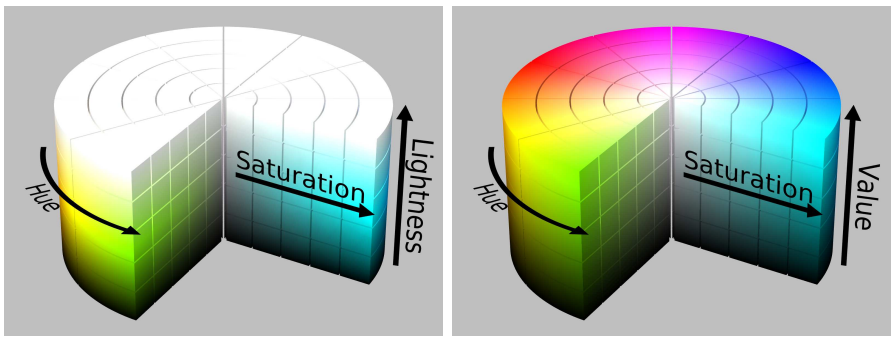


Figure B.2: Cylindrical representation of the HSL (left) and HSV (right) colourspaces. The angle around the central vertical axis corresponds to hue, the distance from the axis is the saturation and the distance along the axis, the lightness or value.

based on the RGB colour space, their values are directly transformable, by means of the following equations:

$$H = \begin{cases} 0° & \Delta = 0 \\ 60° \times (\frac{G-B}{\Delta} mod 6) & C_{max} = R \\ 60° \times (\frac{B-R}{\Delta} + 2) & C_{max} = G \\ 60° \times (\frac{R-G}{\Delta} + 4) & C_{max} = B \end{cases} \tag{B.1}$$

$$S_{HSL} = \begin{cases} 0 & \Delta = 0 \\ \frac{\Delta}{1-|2L-1|} & \Delta <> 0 \end{cases} \tag{B.2}$$

$$L = (C_{max} + C_{min})/2 \tag{B.3}$$

$$S_{HSV} = \begin{cases} 0 & C_{max} = 0 \\ \frac{\Delta}{C_{max}} & C_{max} \neq 0 \end{cases} \tag{B.4}$$

$$V = C_{max} \tag{B.5}$$

where R, G, B are the channel values (normalised between 0 and 1), C_{max} is the maximum of the three channels, C_{min}, the minimum of the three channels and Δ, the difference between C_{max} and C_{min}.

B.3 L*a*b* colour space

The L*a*b* colour space [120] is able to represent all perceivable colours, exceeding the gamuts of several other colour spaces, such as RGB. To that end, the theoretical representation of L*a*b* is a three-dimensional real number space, so it is able to represent an infinite number of colours. In practice, it is usually mapped to a three-dimensional integer space. The term L*a*b* can refer to different colourspaces, but it is used commonly as an abbreviation of CIEL*a*b* colour space, which is the one used in this work.

L*a*b uses three channels: L describes lightness, and a and b represent the colour opponents green-red and blue-yellow. The range of values is 100 for L*, and it varies depending on the RGB colour system for the other two. For example, for sRGB $a \in [-86.185, 98, 254]$ and $b \in [-107.863, 94.482]$. In any case, colours leaning towards negative values are green and blue in a* and b*, respectively, and colours leaning towards positive values are red and yellow in a* and b*, respectively.

One of the main advantages of L*a*b* is that it is device independent. Moreover, it is closer to the human perception than other colourspaces that are designed taken into account the characteristics of specific devices.

To convert an RGB image to L*a*b*, it should be converted to XYZ first, as there is not a direct equivalence among the colour space [121]. The first step is to apply a certain companding function to each of the channels. There are several options to this, and the selection depends on the RGB colour system. In this work, the following formula was applied:

$$v = \begin{cases} \frac{V}{12.92} & V \leq 0.04045 \\ (\frac{V+0.055}{1.055})^{2.4} & otherwise \end{cases} \tag{B.6}$$

where $v \in r, g, b$ and $V \in R, G, B$. Then, the transformation is applied by multiplying the (r, g, b) values by a matrix M, that depends on the RGB working space. In this work, the sRGB (with reference white D65 [122]) was used, so M is the following matrix:

$$M = \begin{pmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{pmatrix} \tag{B.7}$$

To convert from XYZ to L*a*b*, a reference white (X_r, Y_r, Z_r) must be defined beforehand. Using this reference, a function associated to each coordinate is defined. In this work, as a consequence of using the reference white D65, the values are $(X_r = 95.047, Y_r = 100.000, Z_r = 108.883)$. If $\delta_r \in (x_r, y_r, z_r)$, and $(x_r, y_r, z_r) = (X/X_r, Y/Y_r, Z/Z_r)$ then:

$$f_\delta = \begin{cases} \sqrt[3]{\delta_r} & \delta_r > \epsilon \\ \frac{\kappa\delta_r + 16}{116} & otherwise \end{cases} \tag{B.8}$$

where the CIE standard establishes $\kappa = 903.3$ and $\epsilon = 0.008856$. Then, the transformation to L*a*b* is done as follows:

$$L* = 116 f_y - 16 \tag{B.9}$$

$$a* = 500(f_x - f_y) \tag{B.10}$$

$$b* = 200(f_y - f_z) \tag{B.11}$$

B.4 TSL colour space

TSL is a perceptual colour space that defines three channels: tint, saturation and lightness. The tint is defined depending on the closeness of a stimulus to certain main colours: red, green, blue, yellow and white. It is a similar concept as hue plus white. The saturation represents the colourfulness of a certain stimulus relative to its own brightness. The lightness is the brightness measured as closeness to the white. All the three channels are within the same range.

This colourspace was proposed in [123], and is mainly used for face detection or other applications that involve skin recognition. It has a direct transformation from RGB colourspace, as its objective was to make the RGB values more intuitive. The transformation from RGB to TSL is performed as follows:

$$T = \begin{cases} \frac{1}{2\pi} atan(\frac{r'}{g'} + \frac{1}{4}) & g' > 0 \\ \frac{1}{2\pi} atan(\frac{r'}{g'} + \frac{3}{4}) & g' < 0 \\ 0 & otherwise \end{cases} \tag{B.12}$$

$$S = \sqrt{\frac{9}{5} r'^2 + g'^2} \tag{B.13}$$

$$L = 0.299R + 0.587G + 0.114B \tag{B.14}$$

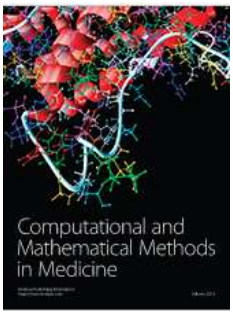
where $r' = \frac{R}{R+G+B} - \frac{1}{3}$ and $g' = \frac{G}{R+G+B} - \frac{1}{3}$.

Appendix C

Publications and other mentions

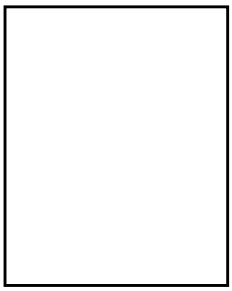
C.1 JCR journals

L. Sanchez, N. Barreira, N. Sanchez-Marono, A. Mosquera, C. Garcia Resua, M.J. Giraldez, On the development of conjunctival hyperemia computer-assisted diagnosis tools: Influence of feature selection and class imbalance in automatic gradings, Artificial Intelligence in Medicine, 71, 30-42, 2016.



L. Sanchez, N. Barreira, A. Mosquera, K. Evans, H. Pena-Verdeal, Defining the optimal region of interest for hyperemia grading in the bulbar conjunctiva, Computational and Mathematical Methods in Medicine, 2016, Article ID 3695014, 1-9, 2016.

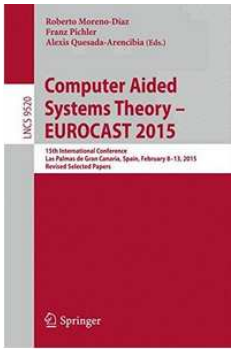
C.2 Book chapters



B. Remeseiro, N. Barreira, L. Sanchez, L. Ramos, A. Mosquera, Machine Learning Applied to Optometry Data, Advances in Biomedical Informatics, 2017.

C.3 Chapters in book series

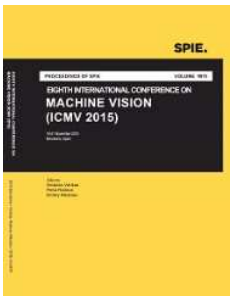
L. Sanchez, N. Barreira, H. Pena Verdeal, E. Yebra Pimentel, A novel framework for hyperemia grading based on artificial neural networks, Lecture Notes in Computer Science: Advances in Computational Intelligence (International Work Conference on Artificial Neural Networks, IWANN 2015), 9094, 263-275, Palma de Mallorca, 2015.



L. Sanchez, N. Barreira, A. Mosquera, C. Garcia Resua, E. Yebra Pimentel, Automatic Selection of Video Frames for Hyperemia Grading, Lecture Notes in Computer Science: Computer Aided Systems Theory, Revised Selected Papers EUROCAST 2015, 9520, 479 - 486, 2015.

C.4 International conferences

L. Sanchez, N. Barreira, A. Mosquera, K. Evans, Assessment of the repeatability in an automatic methodology for hyperemia grading in the bulbar conjunctiva, International Joint Conference on Neural Networks, Anchorage, Alaska, May 2017. CORE A.



L. Sanchez, N. Barreira, N. Sanchez Marono, A. Mosquera, H. Pena Verdeal, E. Yebra Pimentel, On the analysis of local and global features for hyperemia grading, ICMV, Niza, November 2016. CORE C.

L. Sanchez, N. Barreira, N. Sanchez Marono, A. Mosquera, C. Garcia Resua, E. Yebra Pimentel, On the analysis of feature selection techniques in a conjunctival hyperemia grading framework, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 1, 271-276, Bruges, April 2016. CORE B.

L. Sanchez, N. Barreira, A. Mosquera, H. Pena Verdeal, E. Yebra Pimentel, Comparing Machine Learning Techniques in a Hyperemia Grading Framework, International Conference on Agents and Artificial Intelligence (ICAART), 2, 423-429, Roma, February 2016. CORE C.

L. Sanchez, N. Barreira, C. Garcia Resua, E. Yebra Pimentel, Automatic Selection of Video Frames for Hyperemia Grading, Eurocast 2015, 165-166, Las Palmas, Spain, February 2015.

C.5 Under review process

- Precise segmentation of the bulbar conjunctiva for hyperemia images, 2016.
- Automatic Methodology for Hyperemia Grading in the Bulbar Conjunctiva, 2016.

Appendix D

Cohen's kappa

Cohen's kappa [124] is a statistic used to measure inter- or intra-rate agreement. It is more robust than other statistics, such as the correlation coefficient, because it takes into account the accidental agreement, that is, the values that happened by chance, and not due to underlying relationships of the data.

Cohen's kappa measures the agreement between two sources, where each source classifies N items into C categories. The categories must be mutually exclusive. The formulation of κ is the following:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{D.1}$$

where p_0 is the observed agreement among sources, and p_e is the by chance agreement. The latter is computed with the available observed data, and considers the probabilities of each observer randomly observing each category. The formula is the following:

$$p_e = \frac{\sum_k \frac{n_{k1} n_{k2}}{N}}{N} \tag{D.2}$$

where k are the categories, N is the number of items and n_{ki}, the number of items that a source i predicted from a category k.

The kappa index is usually a value between 0 and 1, with 1 meaning a complete agreement and 0, the lowest agreement. Values below zero are unlikely [125] but possible, and interpreted as no agreement at all.

One of the main drawbacks of the kappa index is the difficulty of understanding its results, partially caused by the consideration of the agreement by chance. There are not clear indicators on which values of κ should be considered as high enough. Some authors have provided guidelines, such as [126] or [127]. The former the following ranges: below zero means no agreement; 0-0.20, slight agreement; 0.21-0.40, fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80, substantial agreement; and 0.81-1, almost perfect agreement. The latter defined the following intervals: below 0.40, poor agreement; 0.40-0.75, fair agreement; and over 0.75, excellent agreement. Although both definitions are arbitrary, the five interval one was chosen in this work, in order to provide a thinner classification.

As this statistic is used in categorical items, a transformation of the values has to be made before its application to a continuous range.

Appendix E

Cross-validation

Cross validation is one of the most used validation techniques in machine learning. It is mainly used in the systems where the goal is to predict a result, in order to estimate the accuracy that the model will have in practice.

When working with regression techniques or classifiers, it is a mistake to input the same data to train and to test the system. The model would just learn these patterns, which will lead the system to overfit, that is, to react badly to new inputs that it has not seen before. Therefore, when a machine learning technique is being trained, the dataset is divided in two subsets: training and test. The first part represents the data that the system *studies* in order to learn, while the second part are unknown values that the system comes across for the first time after trained. However, for this approach to succeed, it must be a minimum number of samples available.

Thus, in smaller datasets, other alternatives are used. One of the most well-known is cross-validation [128]. The process of cross-validation divides the available dataset in a number of subsets. Then, it performs an iterative training process, alternating which subsets are used to train and to test. This way, one of the best benefits of cross-validation techniques is that they behave adequately in small datasets, as they use all the data as training and as validation. Moreover, cross-validation provides a more realistic accuracy measure on the results of the system. To reduce the variability, several rounds or iterations of cross-validation are usually run. Then, the validation error will be the average among the iterations.

There are several cross-validation techniques, which can be divided in two main groups: exhaustive and non-exhaustive. The former learn and test all the possible combinations of the input data, that is, they use all the possible data divisions. The latter, on the contrary, use only a subset of the possible partitions of data.

The technique that has been used in this work is 10-fold cross-validation, a non-exhaustive approach. The data is divided in 10 subsets, or folds. Then, for each of these folds, the training process is repeated. In each repetition i, $i \in [1, 10]$, the training set consists of all the instances but the ones that belong to the fold i. These unused instances will constitute the validation set. Once all the 10 folds have been trained and tested, a validation error will exist for each input.

During certain experiments of the work, such as the application of regression techniques in the different feature selection sets, the technique leave-one-out (an exhaustive approach), was also tested. As the results were closely related to 10-fold cross-validation, they were not included.

Appendix F

Resumen

El ojo es uno de los órganos más importantes del cuerpo humano ya que, pese a no ser imprescindible para la vida, tiene una innegable influencia en la mayoría de nuestras tareas cotidianas. Por desgracia, es un elemento muy sensible, propenso a sufrir las repercusiones de diversos problemas de salud, tanto de enfermedades propiamente oculares como de dolencias generales. De ese modo, los ojos muestran indicios tempranos de un gran número de problemas y, por tanto, son el sujeto de numerosas técnicas de diagnóstico.

Algunas de las enfermedades oculares más extendidas son el síndrome del ojo seco o la conjuntivitis alérgica. Ambas tienen una gran incidencia en la población mundial, y el número de casos aumenta cada año. Uno de los síntomas que estas enfermedades tienen en común es la aparición de hiperemia en la conjuntiva bulbar. La hiperemia es una condición clínica que se produce cuando los vasos sanguíneos se atascan, lo que provoca que una gran cantidad de sangre se acumule en la zona. Esta acumulación hace que el tejido afectado adquiera un tono rojizo, conocido como eritema. La hiperemia puede aparecer por culpa de procesos normales en el cuerpo, pero también como indicativo de enfermedades. Debido a ello, es necesario que los clínicos evalúen el nivel de hiperemia que presenta el paciente, con el objetivo de establecer si se encuentra en niveles razonables.

F.1 Evaluación de la hiperemia en la conjuntiva bulbar

La evaluación del nivel de hiperemia conlleva un proceso largo y tedioso. Aunque la evaluación puede hacerse mediante observación directa del ojo, es común grabar un vídeo o sacar fotos, con el objetivo de poder analizarlo cómodamente más tarde. Una vez que se ha grabado un vídeo del ojo del paciente, el optometrista selecciona la imagen que ofrece la mejor representación de la conjuntiva. Dado que los vídeos típicamente duran varios segundos, este proceso puede ser muy largo. A continuación, la imagen seleccionada es analizada. El especialista se fija en una serie de características, tales como el nivel de rojo en la esclerótica o el número de vasos. Estas características se usan para comparar el ojo del paciente con una escala de medida determinada.

Las escalas de evaluación de hiperemia bulbar no son más que colecciones de imágenes, dibujos o fotografías, que muestran los distintos niveles de severidad que presenta el parámetro. Típicamente constan de unos 4 o 5 prototipos, y es el especialista el que infiere los valores intermedios. Para ello, se selecciona el prototipo de la escala que más se asemeja al ojo del paciente y, a continuación, se asigna una parte decimal representando cómo de cerca o lejos están paciente y prototipo. Por tanto, las evaluaciones de hiperemia bulbar se pueden ver como un rango continuo de valores más que como clases separadas.

Además de ser intensivo en tiempo, este proceso manual tiene varios problemas. Por una parte, la subjetividad está presente en todos los pasos, por lo que tanto si se comparan los valores de distintos expertos, como los de un mismo experto en distintos instantes temporales, es común que aparezcan grandes discrepancias entre las evaluaciones. Por otra parte está la dificultad de extraer el conocimiento experto. Cuando un clínico evalúa a un paciente, lo hace basándose en sus experiencias pasadas, y no siempre es enteramente consciente de a qué le está prestando más atención. Por todo ello, es necesario automatizar el proceso.

Este trabajo tiene como objetivo el desarrollo de una metodología automática del cálculo de la hiperemia bulbar que sirva a los expertos como herramienta de ayuda al diagnóstico, facilitando su trabajo y reduciendo el tiempo invertido en una tarea tediosa.

Además, la metodología automática extraerá conocimiento de los datos, permitiendo entender mejor la hiperemia y la forma en la que se realiza su evaluación.

F.2 Metodología

Para el desarrollo de la metodología automática se ha contado con dos bases de datos. Una de ellas, formada por vídeos, ha sido obtenida en el Servicio de Optometría (Universidad de Santiago de Compostela). La otra, compuesta de imágenes, ha sido obtenida en la School of Optometry and Vision Sciences (Cardiff University). Ambas muestran vistas laterales del ojo, donde aparece desde el centro de la pupila hasta el lagrimal o el rabillo del ojo. Con respecto a las escalas de medida, se han usado dos de las más ampliamente aceptadas: la escala Efron, que cuenta con 5 dibujos como prototipos (valores del 0 al 4) y la escala CCLRU, una escala fotográfica con 4 prototipos (valores del 1 al 4).

El sistema recibe un vídeo como entrada. El primer paso es un método de selección automática de la mejor imagen de la secuencia. Hay una serie de condiciones que una imagen debe cumplir para ser considerada adecuada para la evaluación. La iluminación es uno de los parámetros que más influyen en la calidad de una imagen en este contexto, por lo que debe ser suficiente. Además, se debe tener en cuenta que el vídeo muestra un ojo en movimiento, por lo que es necesario detectar y excluir las imágenes borrosas.

Una vez que se ha decidido la mejor imagen teniendo en cuenta iluminación y borrosidad, es necesario delimitar la zona en la que se va a trabajar. Esto es, se realiza un proceso de segmentación automática de la conjuntiva bulbar, eliminando las regiones innecesarias que aparecen en la imagen, tales como los párpados o las pestañas. Estas regiones no contribuyen a la evaluación de la hiperemia y pueden, sin embargo, introducir ruido en los cálculos posteriores. Para separar la conjuntiva de las regiones colindantes, se han estudiado varias aproximaciones con técnicas de procesado de imagen, algunas de ellas adaptadas del estado del arte, como umbralización o detección de contornos, y otras nuevas, desarrolladas para este entorno. Además, se han estudiado técnicas de pre- y post-procesado, con la intención de mejorar los resultados.

Con la conjuntiva adecuadamente delimitada en la imagen, la siguiente etapa consiste en calcular ciertas características de la región que influyen en el nivel de hiperemia. En este trabajo se ha trabajado con un total de 25 características, que calculan valores como el nivel de amarillo en la esclerótica o el grosor medio de los vasos. Las características se han obtenido de trabajos previos del estado del arte, así como de sugerencias de los optometristas. Debido a las diferencias al representar los colores en un medio gráfico, se han estudiado diferentes espacios de color. Además, dado que hay razones para pensar que las distintas regiones del ojo tienen distinta influencia en el nivel de hiperemia, las características se han calculado en tres regiones: en toda la conjuntiva y en cada una de sus mitades, izquierda y derecha. Al calcular parámetros similares y al combinar características de distintas partes es natural que surja una cierta redundancia. Se han utilizado técnicas de selección de características para obtener subconjuntos de características que aportan información significativa. De este modo, se consigue un conjunto mínimo al tiempo que se consigue evitar la pérdida de información.

Tras la determinación del subconjunto de características que se va a utilizar, el último paso de la metodología es utilizar dichas características para calcular el valor final en la escala de medida escogida. Para ello, se utilizarán técnicas de aprendizaje máquina. Dadas las especiales características de los datos, hay dos grandes enfoques que se pueden dar. Por una parte, al ser las escalas colecciones de prototipos finitos, se pueden utilizar clasificadores para resolver el problema. Por otra parte, las escalas son usadas en la práctica como un rango de valores continuos, por lo que se podrían aplicar técnicas de regresión. Ya que cada aproximación tiene pros y contras, se decidió implementar ambas y compararlas. Los resultados obtenidos señalan que los métodos de regresión se adaptan mejor al problema. Además, se realizaron pruebas con características globales y locales, así como la validación con un segundo conjunto de imágenes.

Tras los pasos que han sido descritos hasta este punto, la metodología automática estaba completa y lista para su implantación. Por ello, se procedió a realizar pruebas enfrentándola a algunos de los contratiempos más habituales que aparecen en la práctica.

Uno de los problemas más comunes en entornos de análisis de imagen médica es que no existe un procedimiento estándar de captura de imágenes y vídeos. Esto tiene repercusiones a la hora de conseguir una metodología que generalice bien, ya que las imágenes de una base de datos pueden presentar grandes diferencias con las de otra. Por lo tanto, el primer paso para asegurar que una metodología puede ser implantada en un nuevo entorno es realizar un análisis de repetitibilidad. Así, se ha estudiado cómo reacciona cada paso de la metodología ante variaciones que no afectan a la evaluación de un experto.

Otra preocupación común en este tipo de entornos es el tamaño y rango de la base de datos. Así, es habitual no sólo que haya pocas imágenes disponibles, sino que muchas de las situaciones posibles no están representadas. En el contexto de la hiperemia bulbar, este es un problema frecuente, ya que las escalas de medida no están balanceadas: hay muchos pacientes que son evaluados en los niveles intermedios, pero muy pocos en los extremos. Incluso los individuos sanos suelen presentar al menos trazas de hiperemia, y los casos más graves no suelen verse en atención primaria. Por todo ello, se ha explorado la opción de utilizar técnicas de balanceo de datos para mejorar las características de las bases de imágenes y vídeos disponibles.

Por último, se ha estudiado la importancia de una segmentación precisa. Realizar una segmentación automática de la conjuntiva es una tarea especialmente compleja por culpa de la gran variabilidad de las imágenes, incluso dentro de la misma base de datos. La distancia del ojo a la cámara, la iluminación, la apertura del ojo o la posición del mismo en la imagen son sólo algunos de los parámetros a tener en cuenta. Por ello, incluso habiendo encontrado un método capaz de generalizar hasta cierto punto, nada garantiza que la metodología sea capaz de segmentar de forma óptima nuevos conjuntos de imágenes igualmente válidos. Por lo tanto, se decidió analizar cómo de importante era que la segmentación se adaptase perfectamente al contorno de la conjuntiva. Para esto, se definió una pequeña región central, común a todas las imágenes. Las 25 características se calcularon en esta región central para comprobar su influencia en el cálculo. Además, se realizaron subdivisiones adicionales para obtener pruebas empíricas acerca de qué áreas son las más importantes para los especialistas.

F.3 Resultados

En la selección automática de la mejor imagen de la secuencia de vídeo, considerando los puntos principales de iluminación y borrosidad, la metodología es capaz de obtener una representación adecuada en el 98% de los casos. Además, es capaz de obtener la mejor imagen de la secuencia en un 90% de los casos.

En cuanto a la segmentación de la conjuntiva bulbar, no existe ningún método de segmentación que pueda considerarse universal debido a la variabilidad de los conjuntos de imágenes. Por este motivo, se probó a combinar la salida de varios algoritmos de segmentación. De este modo, se considera como conjuntiva todos aquellos puntos de la imagen que fueran marcados como conjuntiva por al menos 8 algoritmos de segmentación. Se obtiene una precisión, sensitividad y especificidad por encima de 0.8. Debido a la dificultad de extraer con precisión la zona cercana a las pestañas, podemos considerar como válidas las imágenes con un valor de precisión próximo a 0.9. En cuanto a las técnicas de pre- y post-procesado, muchas no ofrecen beneficios que justifiquen su aplicación. Sin embargo, la eliminación de puntos brillantes de las imágenes es aconsejable, ya que estos puntos son información perdida que puede afectar negativamente en la siguiente etapa.

En cuanto a las características de la imagen calculadas en la conjuntiva, aunque hay diferencias entre los conjuntos de imágenes utilizados, el nivel de rojo tanto en la imagen completa como en la esclerótica son las características más repetidas. No hay, sin embargo, un espacio de color que se imponga sobre los demás, ni el nivel de rojo es suficiente, por sí solo, para definir la hiperemia.

Con respecto al último paso de la metodología, al comparar clasificadores y métodos de regresión, estos últimos obtuvieron los mejores resultados. Para comparar sistemas se escogió el error cuadrático medio. Hay que tener en cuenta que es común que los expertos difieran en sus observaciones hasta en 0.5 puntos en las escalas utilizadas, por lo que un error cuadrático medio de 0.25 sería el valor máximo permitido. En este trabajo, utilizando sólo características globales se han obtenido valores de error cuadrático medio de 0.048 y 0.041, en Efron y CCLRU respectivamente. Para el experimento que incluye características locales, los valores obtenidos fueron 0.058 y 0.046, en Efron y

CCLRU respectivamente. Por lo tanto, los valores de la metodología automática están muy por debajo del umbral establecido.

En cuanto a las pruebas realizadas con el objetivo de analizar las posibles dificultades de aplicar la metodología en un entorno real, el análisis de repetitibilidad produjo buenos resultados. En concreto, se evaluó el efecto que dos alteraciones comunes en las imágenes, llevar lentillas y la presencia de restos de una solución de limpieza azulada, tenía en los resultados de cada uno de los pasos. La variabilidad de las salidas es similar a la del especialista, puesto que la diferencia media en evaluaciones del mismo paciente en revisiones consecutivas es de 0.07 en el caso manual y 0.03 en el automático.

La aplicación de técnicas de balanceo también probó su utilidad, dado que es capaz de reducir el error cuadrático medio de un sistema hasta en un 80%. En este caso, las técnicas más adecuadas son las que añaden instancias, dado que los conjuntos de imágenes son reducidos.

Por último, el análisis de un área reducida de medida obtuvo resultados que hacen pensar que un pequeño rectángulo centrado en la imagen es suficiente para obtener una evaluación consistente. Si bien los métodos de selección de características prefieren las características calculadas en toda la conjuntiva, la utilización sólo de características del área reducida puede obtener un error inferior al 0.2, lo que aún está bajo el umbral mínimo permitido. Por lo tanto, incluso considerando sólo una parte de la imagen, el sistema puede imitar los resultados del experto humano.

F.4 Conclusiones

El objetivo de este trabajo era el desarrollo de una metodología automática que evaluara la hiperemia bulbar, capaz de ayudar a los expertos en el diagnóstico eliminando los problemas de la aproximación manual. Además, se buscaba entender mejor el conocimiento que los expertos utilizan para evaluar. Para ello, se ha desarrollado un sistema que recibe un vídeo como entrada, selecciona la mejor imagen de la secuencia, calcula una serie de características relevantes en la región de la conjuntiva y, finalmente, las combina para producir una evaluación en el rango de la escala de medida escogida.

Los resultados muestran que se han cumplido los objetivos, ya que la metodología

puede imitar el comportamiento de los expertos humanos. Además, se ha probado la eficacia de la metodología ante problemas comunes en entornos clínicos, donde también se han obtenido buenos resultados.

Pese a ello, hay varias líneas de investigación que podrían perseguirse a partir de este punto. Una de ellas es la publicación de una base de datos de imágenes, dado que no existen bases de datos públicas en este área. Además, la metodología podría extenderse en diferentes direcciones, como el considerar el seguimiento de un paciente a lo largo de varias revisiones o desarrollar aplicaciones de auto-diagnóstico para usuarios finales.

Bibliography

[1] Cronau, H., Kankanala, R. R., and Mauger, T., "Diagnosis and management of red eye in primary care," *Am Fam Physician*, vol. 81, no. 2, pp. 137–144, 2010.

[2] Wolffsohn, J. S. and Purslow, C., "Clinical monitoring of ocular physiology using digital image analysis," *Contact Lens and Anterior Eye*, vol. 26, no. 1, pp. 27–35, 2003.

[3] Rolando, M. and Zierhut, M., "The ocular surface and tear film and their dysfunction in dry eye disease," *Survey of ophthalmology*, vol. 45, pp. S203–S210, 2001.

[4] Miljanović, B., Dana, R., Sullivan, D. A., and Schaumberg, D. A., "Impact of dry eye syndrome on vision-related quality of life," *American journal of ophthalmology*, vol. 143, no. 3, pp. 409–415, 2007.

[5] Efron, N., Morgan, P. B., and Katsara, S. S., "Validation of grading scales for contact lens complications," *Ophthalmic and Physiological Optics*, vol. 21, no. 1, pp. 17–29, 2001.

[6] Stewart, W. C., Kolker, A. E., Stewart, J. A., Leech, J., and Jackson, A. L., "Conjunctival hyperemia in healthy subjects after short-term dosing with latanoprost, bimatoprost, and travoprost," *American journal of ophthalmology*, vol. 135, no. 3, pp. 314–320, 2003.

[7] Murphy, P. J., Lau, J. S. C., Sim, M. M. L., and Woods, R. L., "How red is a white eye? clinical grading of normal conjunctival hyperaemia," *Eye*, vol. 21, no. 5, pp. 633 – 638, 2006.

[8] Pult, H., Purslow, C., and Murphy, P. J., "The relationship between clinical signs and dry eye symptoms," *Eye*, vol. 25, no. 4, pp. 502–510, 2011.

[9] Ibrahim, O. M., Dogru, M., Takano, Y., Satake, Y., Wakamatsu, T. H., Fukagawa, K., Tsubota, K., and Fujishima, H., "Application of visante optical coherence tomography tear meniscus height measurement in the diagnosis of dry eye disease," *Ophthalmology*, vol. 117, no. 10, pp. 1923–1929, 2010.

[10] Korb, D. R., Herman, J. P., Greiner, J. V., Scaffidi, R. C., Finnemore, V. M., Exford, J. M., Blackie, C. A., and Douglass, T., "Lid wiper epitheliopathy and dry eye symptoms," *Eye & Contact Lens*, vol. 31, no. 1, pp. 2–8, 2005.

[11] Leung, E. W., Medeiros, F. A., and Weinreb, R. N., "Prevalence of ocular surface disease in glaucoma patients," *Journal of glaucoma*, vol. 17, no. 5, pp. 350–355, 2008.

[12] Ohashi, Y., Ebihara, N., Fujishima, H., Fukushima, A., Kumagai, N., Nakagawa, Y., Namba, K., Okamoto, S., Shoji, J., Takamura, E., *et al.*, "A randomized, placebo-controlled clinical trial of tacrolimus ophthalmic suspension 0.1% in severe allergic conjunctivitis," *Journal of Ocular Pharmacology and Therapeutics*, vol. 26, no. 2, pp. 165–174, 2010.

[13] Baudouin, C., Barton, K., Cucherat, M., and Traverso, C., "The measurement of bulbar hyperemia: challenges and pitfalls.," *European journal of ophthalmology*, vol. 25, no. 4, pp. 273–279, 2014.

[14] Fieguth, P. and Simpson, T., "Automated measurement of bulbar redness," *Investigative Ophthalmology and Visual Science*, vol. 43, no. 2, pp. 340–347, 2002.

[15] Schulze, M. M., Jones, D. A., and Simpson, T. L., "The development of validated bulbar redness grading scales," *Optometry & Vision Science*, vol. 84, no. 10, pp. 976–983, 2007.

[16] Efron, N., *Contact lens practice*. Elsevier Health Sciences, 2010.

[17] Yoneda, T., Sumi, T., Takahashi, A., Hoshikawa, Y., Kobayashi, M., and Fukushima, A., "Automated hyperemia analysis software: reliability and reproducibility in healthy subjects," *Japanese journal of ophthalmology*, vol. 56, no. 1, pp. 1–7, 2012.

[18] Rodriguez, J. D., Johnston, P. R., Ousler III, G. W., Smith, L. M., and Abelson, M. B., "Automated grading system for evaluation of ocular redness associated with dry eye," *Clinical ophthalmology (Auckland, NZ)*, vol. 7, p. 1197, 2013.

[19] Peterson, R. C. and Wolffsohn, J. S., "Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia," *British journal of ophthalmology*, vol. 91, no. 11, pp. 1464–1466, 2007.

[20] Wu, S., Hong, J., Tian, L., Cui, X., Sun, X., and Xu, J., "Assessment of bulbar redness with a newly developed keratograph," *Optometry & Vision Science*, vol. 92, no. 8, pp. 892–899, 2015.

[21] Tort, M., Ornberg, R., Lay, B., Danno, R., Soong, F., and Salapatek, A., "Development of an Objective Assessment of Conjunctival Hyperemia Elicited via Conjunctival Allergen Provocation Testing (CAPT) and Environmental Exposure Chamber (EEC) Testing," *EEC (N= 13)*, vol. 2, p. 5, 2012.

[22] Wald, M. J., Lay, B., Danno, R., Grosskreutz, C. L., and Chandra, S., "Performance of automated hyperemia assessment in allergic conjunctivitis interventional study," in *Investigative Ophthalmology & Visual Science*, vol. 56, 2015.

[23] Downie, L. E., Keller, P. R., and Vingrys, A. J., "Assessing ocular bulbar redness: a comparison of methods," *Ophthalmic and Physiological Optics*, vol. 36, no. 2, pp. 132–139, 2016.

[24] Amparo, F., Wang, H., Emami-Naeini, P., Karimian, P., and Dana, R., "The Ocular Redness Index: A Novel Automated Method for Measuring Ocular InjectionA Novel Automated System to Measure Redness," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 7, pp. 4821–4826, 2013.

[25] Wolf, W., "Key frame selection by motion analysis," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, pp. 1228–1231, May 1996.

[26] Erol, B. and Kossentini, F., "Automatic key video object plane selection using the shape information in the mpeg-4 compressed domain," *Multimedia, IEEE Transactions on*, vol. 2, pp. 129–138, Jun 2000.

[27] Radu, P., Ferryman, J., and Wild, P., "A robust sclera segmentation algorithm," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pp. 1–6, IEEE, 2015.

[28] Liu, X., Bowyer, K. W., and Flynn, P. J., "Experiments with an improved iris segmentation algorithm," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pp. 118–123, IEEE, 2005.

[29] van Ginkel, M., Hendriks, C. L., and van Vliet, L. J., "A short introduction to the radon and hough transforms and how they relate to each other," *Delft University of Technology*, 2004.

[30] Kong, W. and Zhang, D., "Accurate iris segmentation based on novel reflection and eyelash detection model," in *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pp. 263–266, IEEE, 2001.

[31] Min, T.-H. and Park, R.-H., "Eyelid and eyelash detection method in the normalized iris image using the parabolic hough model and otsu's thresholding method," *Pattern recognition letters*, vol. 30, no. 12, pp. 1138–1143, 2009.

[32] Mirza, D., Taj, I., Khalid, A., *et al.*, "A robust eyelid and eyelash removal method and a local binarization based feature extraction technique for iris recognition system," in *Multitopic Conference, 2009. INMIC 2009. IEEE 13th International*, pp. 1–6, IEEE, 2009.

[33] Joshi, N., Shah, C., and Kaul, K., "A novel approach implementation of eyelid detection in biometric applications," in *Engineering (NUiCONE), 2012 Nirma University International Conference on*, pp. 1–6, IEEE, 2012.

[34] Ramos, L., Barreira, N., Pena-Verdeal, H., Giráldez, M., and Yebra-Pimentel, E., "Computational approach for tear film assessment based on break-up dynamics," *Biosystems Engineering*, vol. 138, pp. 90–103, 2015.

[35] Remeseiro, B., Oliver, K. M., Tomlinson, A., Martin, E., Barreira, N., and Mosquera, A., "Automatic grading system for human tear films," *Pattern Analysis and Applications*, vol. 18, no. 3, pp. 677–694, 2015.

[36] Peterson, R. C. and Wolffsohn, J. S., "Objective grading of the anterior eye," *Optometry & Vision Science*, vol. 86, no. 3, 2009.

[37] Papas, E. B., "Key factors in the subjective and objective assessment of conjunctival erythema," *Investigative ophthalmology & visual science*, vol. 41, no. 3, pp. 687–691, 2000.

[38] Park, I. K., Chun, Y. S., Kim, K. G., Yang, H. K., and Hwang, J.-M., "New clinical grading scales and objective measurement for conjunctival injection," *Investigative ophthalmology & visual science*, vol. 54, no. 8, pp. 5249–5257, 2013.

[39] Bailey, I. L., Bullimore, M. A., Raasch, T. W., and R., T. H., "Clinical grading and the effects of scaling," *Investigative ophthalmology & visual science*, vol. 32, no. 2, pp. 422–32, 1991.

[40] Al-Tairi, Z. H., Rahmat, R. W. O., Saripan, M. I., and Sulaiman, P. S., "Skin segmentation using yuv and rgb color spaces.," *JIPS*, vol. 10, no. 2, pp. 283–299, 2014.

[41] Sun, Y., Duthaler, S., and Nelson, B. J., "Autofocusing algorithm selection in computer microscopy," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 70–76, IEEE, 2005.

[42] Fairchild, M. D., *Color appearance models*. John Wiley & Sons, 2013.

[43] Al-Amri, S. S., Kalyankar, N. V., *et al.*, "Image segmentation by using threshold techniques," *arXiv preprint arXiv:1005.4020*, 2010.

[44] Kakumanu, P., Makrogiannis, S., and Bourbakis, N., "A survey of skin-color modeling and detection methods," *Pattern recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[45] Sigal, L., Sclaroff, S., and Athitsos, V., "Skin color-based video segmentation under time-varying illumination," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 7, pp. 862–877, 2004.

[46] Mo, S., Cheng, S., and Xing, X., "Hand gesture segmentation based on improved kalman filter and tsl skin color model," in *Multimedia Technology (ICMT), 2011 International Conference on*, pp. 3543–3546, 2011.

[47] Suzuki, S. *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[48] Lee, J. S., Haralick, R. M., and Shapiro, L. G., "Morphologic edge detection," *Robotics and Automation, IEEE Journal of*, vol. 3, no. 2, pp. 142–156, 1987.

[49] Yang, T., Yang, L.-B., Wu, C. W., and Chua, L. O., "Fuzzy cellular neural networks: applications," in *Cellular Neural Networks and their Applications, 1996. CNNA-96. Proceedings., 1996 Fourth IEEE International Workshop on*, pp. 225–230, IEEE, 1996.

[50] Cousty, J., Bertrand, G., Najman, L., and Couprie, M., "Watershed cuts: Minimum spanning forests and the drop of water principle," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 8, pp. 1362–1374, 2009.

[51] Beucher, S. and Meyer, F., "The morphological approach to segmentation: the watershed transformation," *OPTICAL ENGINEERING-NEW YORK-MARCEL DEKKER INCORPORATED-*, vol. 34, pp. 433–433, 1992.

[52] Beucher, S. and Lantuéjoul, C., "Use of watersheds in contour detection," in *International workshop on image processing, real-time edge and motion detection*, 1979.

[53] Wu, X., "Adaptive split-and-merge segmentation based on piecewise least-square approximation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 8, pp. 808–815, 1993.

[54] Horowitz, S. L. and Pavlidis, T., "Picture segmentation by a tree traversal algorithm," *Journal of the ACM (JACM)*, vol. 23, no. 2, pp. 368–388, 1976.

[55] Rivero, J. S. and Bouthemy, P., "Region segmentation according to motion-based criteria," 1987.

[56] Deng, G. and Cahill, L., "An adaptive gaussian filter for noise reduction and edge detection," in *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pp. 1615–1619, IEEE, 1993.

[57] Wang, Z. and Zhang, D., "Progressive switching median filter for the removal of impulse noise from highly corrupted images," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 46, no. 1, pp. 78–80, 1999.

[58] Tomasi, C. and Manduchi, R., "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*, pp. 839–846, IEEE, 1998.

[59] Stanković, R. S. and Falkowski, B. J., "The haar wavelet transform: its status and achievements," *Computers & Electrical Engineering*, vol. 29, no. 1, pp. 25–44, 2003.

[60] Funt, B., Barnard, K., and Martin, L., "Is machine colour constancy good enough?," in *Computer Vision ECCV'98*, pp. 445–459, 1998.

[61] Finlayson, G. D., Drew, M. S., and Funt, B. V., "Color constancy: generalized diagonal transforms suffice," *JOSA A*, vol. 11, no. 11, pp. 3011–3019, 1994.

[62] Van De Weijer, J., Gevers, T., and Gijsenij, A., "Edge-based color constancy," *IEEE Transactions on image processing*, vol. 16, no. 9, pp. 2207–2214, 2007.

[63] Van De Weijer, J. and Gevers, T., "Color constancy based on the grey-edge hypothesis," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2, pp. II–722, IEEE, 2005.

[64] Barnard, K., Cardei, V., and Funt, B., "A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data," *Image Processing, IEEE Transactions on*, vol. 11, no. 9, pp. 972–984, 2002.

[65] Rizzi, A., Gatta, C., and Marini, D., "Color correction between gray world and white patch," in *Electronic Imaging 2002*, pp. 367–375, International Society for Optics and Photonics, 2002.

[66] Canny, J., "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 679–698, 1986.

[67] Vázquez, S., Barreira, N., Penedo, M. G., Pena-Seijo, M., and Gómez-Ulla, F., "Evaluation of SIRIUS retinal vessel width measurement in REVIEW dataset," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, June 20-22, 2013* (Rodrigues, P. P., Pechenizkiy, M., Gama, J., Cruz-Correia, R., Liu, J., Traina, A. J. M., Lucas, P. J. F., and Soda, P., eds.), pp. 71–76, IEEE Computer Society. Washington, DC, 2013.

[68] Holland, P. W. and Welsch, R. E., "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[69] Jolliffe, I., *Principal component analysis*. Wiley Online Library, 2002.

[70] Ye, J., Janardan, R., Li, Q., *et al.*, "Two-dimensional linear discriminant analysis.," in *NIPS*, vol. 4, p. 4, 2004.

[71] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A., *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.

[72] Gennari, J. H., Langley, P., and Fisher, D., "Models of incremental concept formation," *Artificial intelligence*, vol. 40, no. 1-3, pp. 11–61, 1989.

[73] Das, S., "Filters, wrappers and a boosting-based hybrid for feature selection," in *ICML*, vol. 1, pp. 74–81, 2001.

[74] Hall, M. and Smith, L., "Practical feature subset selection for machine learning," in *Australian Computer Science Conference*, pp. 181–191, 1998.

[75] Fayyad, U. M. and Irani, K. B., "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993* (Bajcsy, R., ed.), pp. 1022–1029, Morgan Kaufmann. Burlington, Massachusetts, 1993.

[76] Robnik-Šikonja, M. and Kononenko, I., "An adaptation of Relief for attribute estimation in regression," in *Machine Learning: Proceedings of the Fourteenth International Conference*, pp. 296–304, 1997.

[77] Kira, K. and Rendell, L. A., "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence. San Jose, CA, July 12-16, 1992* (Swartout, W. R., ed.), vol. 2, pp. 129–134, AAAI Press / The MIT Press. Palo Alto, California, 1992.

[78] Holmes, G., Hall, M., and Prank, E., *Generating rule sets from model trees*. Springer, 1999.

[79] Quinlan, J. R. *et al.*, "Learning with continuous classes," in *Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16-18 November 1992* (Adams, A. and Sterling, L., eds.), vol. 92, pp. 343–348, World Scientific Pub Co Inc. Singapore, 1992.

[80] Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K., "Improvements to the smo algorithm for svm regression," *Neural Networks, IEEE Transactions on*, vol. 11, no. 5, pp. 1188–1193, 2000.

[81] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[82] Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[83] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., "Classification and regression trees. wadsworth & brooks," *Monterey, CA*, 1984.

[84] Aha, D. W., Kibler, D., and Albert, M. K., "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[85] Dudani, S. A., "The distance-weighted k-nearest-neighbor rule," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 325–327, 1976.

[86] Kohonen, T., "Improved versions of learning vector quantization," in *International Joint Conference on Neural Networks*, pp. 545–550, 1990.

[87] Baum, E. B., "On the capabilities of multilayer perceptrons," *Journal of complexity*, vol. 4, no. 3, pp. 193–215, 1988.

[88] John, G. H. and Langley, P., "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[89] Abdi, H., "Partial least square regression (PLS regression)," *Encyclopedia for research methods for the social sciences*, pp. 792–795, 2003.

[90] Park, J. and Sandberg, I. W., "Universal approximation using radial-basis-function networks," *Neural computation*, vol. 3, no. 2, pp. 246–257, 1991.

[91] Breiman, L., "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[92] Kohonen, T., "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[93] Smola, A. J. and Schölkopf, B., "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[94] Wang, Y. and Witten, I. H., "Induction of model trees for predicting continuous classes," 1996.

[95] Rousseeuw, P. J. and Leroy, A. M., *Robust regression and outlier detection*. 1987.

[96] Friedman, N., Geiger, D., and Goldszmidt, M., "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[97] Jensen, F. V., *An introduction to Bayesian networks*, vol. 210. UCL press London, 1996.

[98] Kohavi, R., "The power of decision tables," in *Machine Learning: ECML-95*, pp. 174–189, Springer, 1995.

[99] Quinlan, J. R., *C4. 5: programs for machine learning*. Elsevier, 2014.

[100] Holte, R. C., "Very simple classification rules perform well on most commonly used datasets," *Machine learning*, vol. 11, no. 1, pp. 63–90, 1993.

[101] Breiman, L., "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[102] Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[103] Platt, J. *et al.*, "Fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods - support vector learning*, vol. 3, 1999.

[104] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K., "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[105] Hastie, T., Tibshirani, R., *et al.*, "Classification by pairwise coupling," *The annals of statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[106] Cooper, G. F. and Herskovits, E., "A bayesian method for constructing bayesian belief networks from databases," in *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pp. 86–94, Morgan Kaufmann Publishers Inc., 1991.

[107] Walker, J., Young, G., Hunt, C., and Henderson, T., "Multi-centre evaluation of two daily disposable contact lenses," *Contact Lens and Anterior Eye*, vol. 30, no. 2, pp. 125–133, 2007.

[108] Wolffsohn, J. S., Hunt, O. A., and Chowdhury, A., "Objective clinical performance of 'comfort-enhanced' daily disposable soft contact lenses," *Contact Lens and Anterior Eye*, vol. 33, no. 2, pp. 88–92, 2010.

[109] Murphy, P., Lau, J., Sim, M., and Woods, R., "How red is a white eye? clinical grading of normal conjunctival hyperaemia," *Eye*, vol. 21, no. 5, pp. 633–638, 2007.

[110] Wolffsohn, J., "Incremental nature of anterior eye grading scales determined by objective image analysis," *British journal of ophthalmology*, vol. 88, no. 11, pp. 1434–1438, 2004.

[111] Peterson, R. C. and Wolffsohn, J. S., "Objective grading of the anterior eye," *Optometry & Vision Science*, vol. 86, no. 3, pp. 273–278, 2009.

[112] Zhou, Z.-H. and Liu, X.-Y., "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 1, pp. 63–77, 2006.

[113] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[114] Bradski, G. *et al.*, "The opencv library," *Doctor Dobbs Journal*, vol. 25, no. 11, pp. 120–126, 2000.

[115] The MathWorks, Inc., *MATLAB and Statistics Toolbox Release 2014a*. The MathWorks, Inc., Natick, Massachusetts, United States, 2014.

[116] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[117] Tkalcic, M. and Tasic, J. F., *Colour spaces: perceptual, historical and applicational background*, vol. 1. IEEE, 2003.

[118] Joblove, G. H. and Greenberg, D., "Color spaces for computer graphics," in *ACM siggraph computer graphics*, vol. 12, pp. 20–25, ACM, 1978.

[119] Smith, A. R., "Color gamut transform pairs," *ACM Siggraph Computer Graphics*, vol. 12, no. 3, pp. 12–19, 1978.

[120] Colorimetry, C., "official recommendations of the international commission on illumination," *Paris: Commission Internationale de l'Èclairage [International Commission on Illumination]*, 1976.

[121] Connolly, C. and Fleiss, T., "A study of efficiency and accuracy in the transformation from rgb to cielab color space," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 1046–1048, 1997.

[122] Noboru, O. and Robertson, A. R., "3.9: Standard and supplementary illuminants, colorimetry," 2005.

[123] Terrillon, J.-C., David, M., and Akamatsu, S., "Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 112–117, IEEE, 1998.

[124] Cohen, J., "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[125] McHugh, M. L., "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[126] Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[127] Fleiss, J. L., Levin, B., and Paik, M. C., *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[128] Kohavi, R. *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Stanford, CA, 1995.